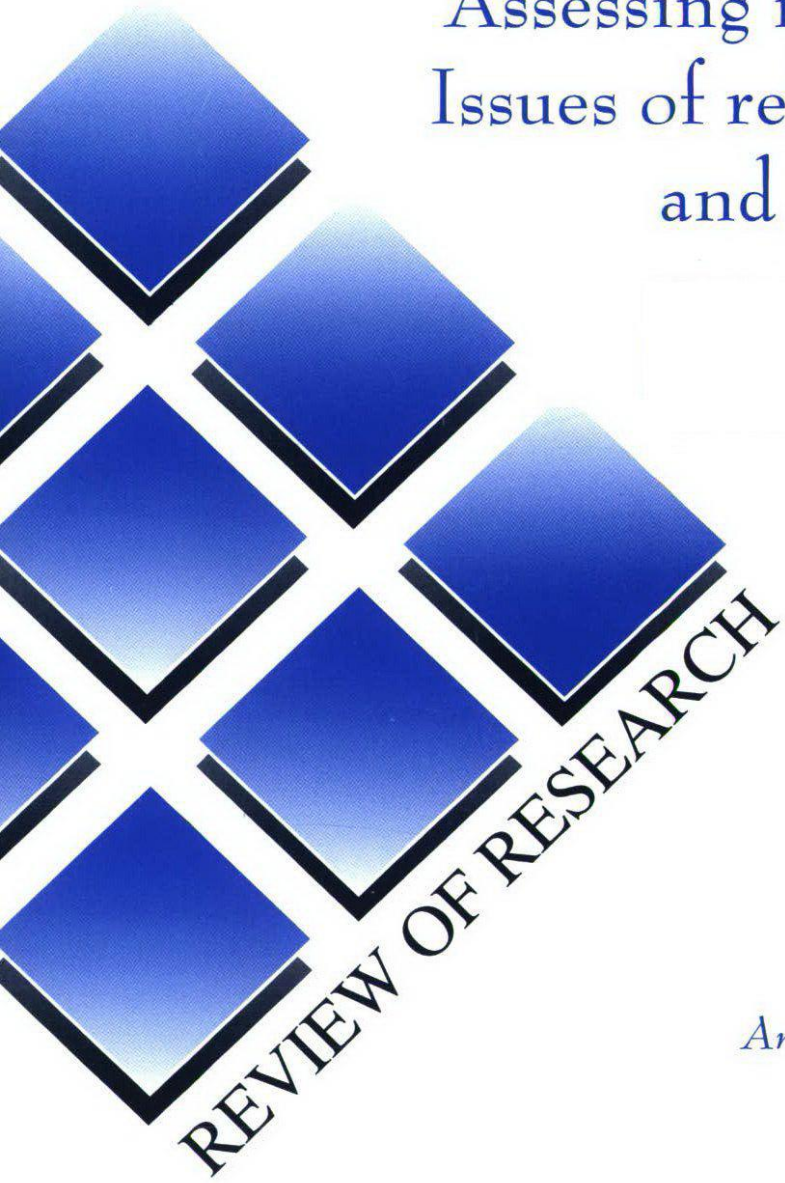
NCVER

Assessing in VET: Issues of reliability and validity



Shelley Gillis
Andrea Bateman

Assessing in VET: Issues of reliability and validity

Shelley Gillis
Andrea Bateman

NCVER

Acknowledgements

This project was carried out under the auspices of the National Research and Evaluation Committee, a sub-committee of the Board of the National Centre for Vocational Education Research. The authors greatly acknowledge the contribution of Professor Patrick Griffin of the Assessment Research Centre at the University of Melbourne who provided invaluable support, technical expertise and guidance.

© Australian National Training Authority, 1999

This work has been produced by the National Centre for Vocational Education Research (NCVER) with the assistance of funding provided by the Australian National Training Authority (ANTA). It is published by NCVER under licence from ANTA. Apart from any use permitted under the Copyright Act 1968, no part of this publication may be reported by any process without the written permission of NCVER Ltd. Requests should be made in writing to NCVER Ltd.

The views and opinions expressed in this document are those of the author/project team and do not necessarily reflect those of the Australian National Training Authority.

ISBN: 0 87397 542 1
TD/TNC: 58.27

Published by NCVER
ACN 007 967 311
252 Kensington Road, Leabrook, SA 5068
PO Box 115, Kensington Park, SA 5068, Australia
www.ncver.edu.au

Contents

Executive summary

Main conclusions

VALIDITY OF AN assessment refers to the use and interpretation of evidence collected, as opposed to the assessment method or task. It is not simply a property of the assessment task. An assessment task that is highly valid for one use or context may be invalid for another.

There are a number of different types of validity, including face, content, construct, criterion (concurrent and predictive) and consequential. Each type needs to be considered when designing assessment tasks and/or interpreting assessment outcomes for a particular purpose.

Validity is largely determined through inferences made by both the task developers and users.

An essential component of the validity of assessments is the assessor's intention. Assessors should be very clear about their intentions when assessing candidates against competency standards, and should identify the boundaries and limitations of the interpretations they make of assessments for a particular purpose and context.

The validity of workplace assessments is often defended on the grounds of the authentic nature of the assessments. Although this provides evidence of face validity, further evidence of content, criterion, construct and consequential validity is needed before the assessment can be said to be valid.

The reliability of an assessment is an estimate of how *accurate* or *precise* the task is as a measurement instrument. Reliability is concerned with how much error is included in the evidence.

There are common sources of error associated with both objective tests and performance assessment. These are associated with:

- ❖ the method of gathering evidence (i.e. the level of precision of the assessment task and the degree of standardisation of the administration and scoring procedures)
- ❖ the characteristics of the candidate (e.g. fatigue if a long test)

In performance assessment, there are additional sources of error:

- ❖ the characteristics of the assessor (e.g. preconceived expectations of the competency level of the candidate)
- ❖ the context of the assessment (e.g. location)
- ❖ the range and complexity of the task(s) (e.g. the level of contextualisation)

Each of the above factors need to be controlled throughout the assessment in order to improve reliability. Assessment procedures need to be developed to minimise the error in the evidence collected and interpreted by assessors. Establishing clear task specifications, including evidence to be collected and decision-making rules, will increase reliability.

Evidence is crucial in establishing reliability and validity of assessments. The methods used to collect the evidence will impact on the reliability, whilst the way in which assessors use and interpret the evidence collected will impact on the validity of the assessment. As reliability creates a foundation for validity, an assessment should aim to reduce the error or 'noise' in the evidence collected or used.

Validation of an assessment process should therefore address the various forms of reliability and validity. It will require the assessment task developers and users (i.e. assessors) to make an holistic judgement as to whether this evidence supports the intended use and interpretation of assessment evidence for the specified purpose(s). The intended use, context and limitations of the assessment task need to be reported to potential users. Ultimately, the validation of an assessment requires evidence of task development, clear and concise assessment criteria against the competency standards, appropriate task administration procedures, adequate scoring/decision-making rules and recording procedures.

Findings and directions for further research

The review of literature has revealed a number of areas requiring further research. These include research into:

- ❖ validation approaches used by workplace assessors and VET practitioners within Australia
- ❖ transferability of competencies outside the assessment event

- consequences of competency-based assessments in both vocational educational settings and the workplace
- factors that influence judgements in competency-based assessment and how such factors impact on reliability and validity

Context

WITH THE IMPLEMENTATION of competency-based assessments within the Australian vocational education and training (VET) system, there is a need to rethink many fundamental concepts of assessment—including the notions of reliability and validity—and to examine procedures for ensuring the development and use of high quality assessment procedures and tasks, which result in valid and reliable judgements of competence.

While much has been written on the theory, policy and practice of competency-based assessment (CBA), very little empirical research on these issues has been identified so far, despite the wide-scale implementation of CBA in Australia, New Zealand, the United Kingdom and Germany. To date, there has been no Australian empirical research into the factors that may influence the reliability and validity of assessment judgements across different contexts (e.g. location and competency domains). There are, however, a number of lessons to be learnt from the substantial body of international studies that have examined reliability and validity issues associated with two major forms of assessment: 'objective' and 'performance assessments'. (For the purposes of this review, the term 'objective' will be used to refer to paper-based objective testing techniques.)

Within a CBA system, the adoption of both forms of assessment helps to ensure that assessments measure both the underpinning knowledge and understanding—as well as the skills required for competent performance in the workplace. The issue is not which form of assessment (i.e. objective versus performance) is more appropriate for use within the VET sector. Rather, it is the appropriateness and importance of the different types of reliability and validity that need to be evaluated according to the purposes of the assessment, and the way in which the evidence will be interpreted and used.

This publication therefore reviews the Australian discussion papers on reliability and validity, as well as the international empirical research in this field.

Introduction

WITH THE IMPLEMENTATION of CBA within the Australian VET system, there is a need to re-evaluate and apply fundamental concepts of assessment, including the notions of reliability and validity. There is also a need to examine methods for ensuring the development and use of high quality assessment procedures and tasks, given that CBA underpins the success of the Australian Recognition Framework (ARF). There are a number of lessons to be learnt from the substantial body of national and international literature, which have reported on the use and evaluation of objective testing methods (such as paper and pencil tests) and, more recently, performance assessment methods. In this paper, we have drawn upon the substantial body of research on classical test theory (typically associated with standardised objective testing) to illustrate how the fundamental principles of reliability and validity can be applied to CBA.

Background information

The Federal Government is committed to the development of a training market to raise the quality, diversity and efficiency of the Australian VET system. Underpinning this training market is the quality of the assessments conducted for recognition purposes.

Since the introduction of the ARF, assessments of competency in vocational educational and workplace settings have become increasingly important. Greater importance attached to assessment methods has been due to greater involvement of industry in the development of competency standards, training packages and recognition. This trend is evident throughout Australia, and within the United Kingdom's and New Zealand's VET systems.

Competency-based assessments can now be conducted for credentialling purposes within the Australian Qualifications Framework (AQF). Hence there is a need to review the appropriateness of a range of assessment

methods that are currently being used. The range of methods tends to be classified into two major forms: 'performance' versus 'objective' (Hayton and Wagner 1998). The former tends to describe assessment methods that require the candidate to actively generate or create a response/product that demonstrates their knowledge or skill (Elliot 1994). Examples include: portfolio, simulations, role-plays, practical demonstrations, workplace observations, open-ended questions, peer/self/supervisor assessments and oral presentations. The latter refers to paper-based objective testing techniques, in which the candidate selects a response from a range of alternatives established by the task developers (e.g. multiple choice, true/false questions). CBA encompasses the use of both objective testing techniques and performance tasks to gather evidence of competence.

CBA, as opposed to curriculum, is now a major quality assurance mechanism in the VET system. Despite the existence of industry training packages, competency standards and training programs for assessors, the selection of assessment methods, the design of assessment tasks and the making of a judgement is still complex and confusing for the assessors. Assessment guidelines, which form part of every training package, require industries to outline preferred assessment methods and to make recommendations for their use. Much of the information included in such guidelines was acquired through direct consultations with industry personnel. This helps to achieve industry acceptance and credibility.

When selecting assessment methods, assessors are currently guided by training packages, and various publications and training resources (e.g. Clayton 1995; Gonczi, Hager and Athanasou 1993; Griffin and Gillis 1997; Hager, Athanasou and Gonczi 1994). For instance, Clayton (1995) provides advice on appropriate methods for assessing skills, knowledge and attitudes (i.e. different domains of competency). Recent research, however, indicates that assessors tend to select assessment methods they have had the most experience and familiarity with. Consequently, continuous exposure tends to reinforce practice (Gillis, Griffin, Trembath and Ling 1997).

However, to date there has been little empirical research into the validation of methods of assessment. In particular, the influence that the method(s) has on the reliability and validity of assessment outcomes across different contexts (e.g. location, competency domains) remains unexamined. Issues such as simplicity, ease of use and cost effectiveness have tended to be the criteria for inclusion of recommended assessment methods in the assessment guidelines of training packages.

As this rapidly changing VET environment moves towards greater accountability, CBA will continue to be a primary task of industry trainers,

workplace assessors and teachers. Under conditions of accountability, the trainers, assessors and teachers will need to understand the impact that different assessment methods have on the reliability and validity of CBA decisions and outcomes.

How assessors select evidence gathering methods for subsequent task development is critical to the success of competency-based training and assessment in the VET system. This paper explores the issues surrounding the selection of evidence-gathering methods, and the ultimate pursuit of reliability and validity in CBA.

What is competency?

Competency comprises the specification of the complex combination of knowledge, skills and attitudes required for successful performance in the workplace (Masters and McCurry 1990). It requires inferences to be made by the assessor as to whether competence is demonstrated. The National Training Board (NTB) policy documents clearly state that the concept of competency was intended to capture broader aspects of work performance, and that the contextual issues need to be considered in this approach (NTB 1992). Work competence includes:

- performance at an acceptable level of technical skill
- organisation of one's tasks
- appropriate response and reaction when things go wrong
- fulfilment of a role in the scheme of things at work
- transfer of skills and knowledge to new situations

What is competency-based assessment?

Prior to exploring any technical issues associated with assessment, we will first clarify what is meant by the term 'competency-based assessment'. Although there are disparate views on the exact definition of assessment—with the terms 'measurement', 'evaluation' and 'testing' often used interchangeably—it can be argued that there are common essential components that comprise an assessment and reporting model. Griffin and Nix (1991) defined assessment as the *purposeful* process of *gathering* appropriate and sufficient *evidence* of competence, and the *interpretation* of that evidence to enable a *judgement*. Included in this model is the *recording* of the evidence and the decision, as well as the *communication* of the outcomes to key *stakeholders*. Therefore at a minimum, CBA should:

- clearly define the *purpose*(s) (e.g. credentialling, promotion, recruitment)

- identify and document the *evidence* required to demonstrate competency
- use appropriate evidence gathering *methods*
- *interpret* the evidence against the competency standards and make a judgement
- establish and use *record keeping* procedures
- *report* appropriate outcomes of the assessment to key stakeholders

This criterion-referenced assessment model encapsulates definitions provided by national and international educational researchers (Athanasou 1997; Clayton 1995; Glaser 1981; Griffin and Gillis 1997; Griffin and Nix 1991; Messick 1992). Each step in the model can help in understanding further issues of reliability and validity, and how they are affected.

Gonczi et al. (1993) argue that the aim of CBA is to assess the attributes underpinning competent performance in the most realistic, holistic and direct way possible. Griffin (1997) goes further, arguing that the major purpose of CBA is the *prediction* of workplace performance. Both views encourage the use of performance tasks to gather evidence of competence. However, objective tests are argued to be useful tools for assessing underpinning knowledge and understanding within a CBA system (Masters and McCurry 1990). Recent research has found that competency-based assessors use both 'objective' and 'performance' forms of assessment in both the classroom and workplace setting, for purposes such as credentialling, needs analysis and remuneration (Gillis et al. 1998; Gillis et al. 1997). The adoption of multiple forms of assessment helps to ensure that assessments measure both the underpinning knowledge and understanding, as well as the skills required for competent performance in the workplace.

What is a reliable and valid assessment?

Although there is consensus that all assessments must be reliable and valid, technical terms such as 'reliability' and 'validity' are often ascribed to assessments without sufficient supporting evidence, and with a range of different meanings. For instance, in a recent investigation of CBA practices in VET in school studies, it was common practice for both teachers and workplace supervisors to argue that their assessments of vocational learning were valid indicators of levels of competence. This was because either the assessment task(s) had been accepted and endorsed by specialists in the field, or the assessments were based on national competency standards and were therefore assumed to be valid (Gillis et al. 1998). Although it is important for assessments to be endorsed by specialists in the field and to match the

competency standards, these two conditions alone do not provide sufficient evidence to conclude the validity of the assessment.

Even among educational researchers who specialise in assessment and measurement, there is disagreement about what constitutes validity. For instance, Hager et al. (1994) argue that validity of an assessment refers to the extent to which the assessment method measures what it is supposed to measure. There are two parts to this definition:

- how well the task measures—concerned with precision of the tools/instruments
- the accurate and clear definition of what it supposes to measure—clarity of the performance criteria and evidence guides within the competency standards

In many assessment instances, both of these tend to be ignored.

Others emphasise the importance of the evidence collected, and the way in which that evidence is interpreted and used for its stated purpose (e.g. Bennett 1993; Linn 1994; Messick 1992; Wilson, Scherbarth, Brickell, Mayo and Paul 1988). This is an equally demanding definition. According to this view, validity of an assessment refers to *use* and *interpretation* of the *evidence* collected, as opposed to the assessment method or task. Consequently, an assessment task that has been developed and validated for one specific purpose and target group, may not necessarily be a valid assessment task for another purpose or target group.

Cronbach (1971) argues that assessors do not validate a task/test, but instead validate the interpretation of the assessment evidence gathered for a specific purpose. For example, an assessment task that has been designed for credentialling purposes within a vocational setting may not be an appropriate task to use when assessing for promotional purposes in an industrial setting—despite the fact that the task has been designed to measure the same unit of competency. Therefore, any assessment task has to be validated in light of the purpose of the assessment, and this will depend upon the accurate interpretation of the evidence collected. As we will see later, the second approach may have important implications for conceptualising reliability and hence validity.

Types of validity

There are several different types of validity that are considered when validating an assessment, with the most widely cited being:

- face
- content

- ❖ criterion related (predictive and concurrent)
- ❖ construct
- ❖ consequential

Each of these validity forms is described in table 1 (on page 12). Although there is widespread agreement that face validity is not a true form of validity in the technical sense (Messick 1989; Wiggins 1991), its importance lies within the acceptance and credibility of the assessment outcomes by the key stakeholders. Given the need for industry to accept the assessment outcomes of the VET sector, the importance of face validity in a CBA system cannot be underestimated.

A unitary notion of validity has been proposed by Messick (1989). He argues that construct validity embraces both content and criterion-related validity, but does not capture the notion of consequential validity. Therefore, validation of an assessment process requires an evaluation of the interpretation of results, as well as the intended and unintended consequences of using the assessment (Elliot 1994). Others such as Zeller (1989) and Cronbach (1984) argue that evidence of each type of validity is required to make an overall judgement of the validity of the assessment, and therefore these validity types should not be treated as alternatives. However, meeting the requirements of one type of validity (e.g. content) is not sufficient to validate an assessment process. Any assessment process needs to address multiple types of validity (Cronbach 1984).

Types of reliability

The reliability of an assessment refers to its degree of stability, consistency and accuracy of the assessment outcomes (Bennett 1993; Groth-Marnat 1990; Kerlinger 1973; Messick 1992). It refers to the extent to which an assessment can theoretically provide repeatable outcomes for candidates of equal competence at different times and/or places. In a sense, reliability is an estimate of how accurate or precise the task is as a measurement instrument.

In classical test theory, reliability is considered to be the relationship between true score and error. In the classical sense, it is assumed that any assessment result is made up of two components: the true ability and error (Thorndike 1988). The reliability is usually expressed as the ratio between the variance of the true ability and the variance of errors of measurement. In simple terms, reliability is about the extent to which error is included in the evidence. Reliability is often reported and interpreted as a statistical concept. It tends to be reported as a correlation co-efficient, particularly when associated with objective testing techniques administered through paper and pencil formats.

The traditional measures of reliability associated with objective testing include 'parallel forms' and 'internal consistency'.

If we think of any assessment as consisting of a judgement or inference, then all judgements or inferences are based upon evidence. Most performances will vary from day to day or from context to context. Judging an individual's competence is a complex task (Guthrie 1993; Block, Clayton and Favero 1995). When performance is to be assessed, competency is inferred from relevant observations of behaviour. Not all the evidence is, or can be, accurately interpreted. There will always be a certain degree of error present in any assessment event. Reliability could be considered as the degree to which the evidence is accurately interpreted. Increasing reliability then becomes a process of controlling or eliminating the factors that reduce the accuracy of interpretation. Estimates of reliability are an indication of how successful that process has been.

The extent to which the interpretation is inaccurate is often called measurement error, but this is a generic term covering all factors that influence the accuracy of the interpretation. Factors that are internal to the candidate include level of fatigue, motivation, interest, nervousness and practice effects (Athanasou 1997; Griffin and Nix 1991; Groth-Marnat 1990; Rudner 1994). Those factors external to the candidate include assessor biases (e.g. attitude toward the candidate), poor administration conditions, adequacy of scoring or coding procedures, or the design of the assessment task itself.

Competency-based assessments are very rarely conducted under ideal conditions. While it could be argued that the workplace context is an ideal setting and relies on direct observation, it may be less than ideal for synthesising, interpreting and evaluating evidence to make an holistic judgement. Assessments are complex events. An assessor needs to attend simultaneously to numerous events, people, circumstances and tasks. In many instances, assessment is only a part of the overall responsibilities of the workplace assessor. In an industrial setting, contextual factors can influence judgements of different individuals performing at the same level of competence. They can alter the interpretation and use of criteria in the judgement process. If those influences can be controlled, then the assessments will be more reliable and will reflect the 'true' competence of the candidate. Assessment procedures need to minimise the influences of confounding sources of evidence.

Table 1: Validity types—Definitions and relevance to CBA

Type of validity	Description	Examples	Determining validity	Importance to CBA
Face	The assessment tasks should be designed to look like they are assessing what they claim to be assessing.	If assessing computing competencies, the assessment task could be designed to collect direct evidence of computing skills through practical demonstrations or simulations. Assessment of the underlying knowledge and computing skills through paper and pencil tests alone would not have face validity.	Judgements of the task users (i.e. those who are going to use the assessment information) requires application of common sense, usually reinforced by expert opinion.	For acceptance of assessment outcomes by key stakeholders (e.g. supervisors, management and candidates); can have political/industrial implications.
Content	Concerned with the extent to which the skills and knowledge demonstrated in the assessment task constitute a representative sample of the skills and knowledge to be exhibited in the competency standards.	When there is a direct match between the required knowledge and skills specified in the standards and the assessment task's capacity to collect such evidence. For example, when the elements of competency are assessed through direct observation of workplace performance, the elements are treated as tasks to be demonstrated.	Requires judgements and inferences to be made by the task developers as to whether the content domain of the task (i.e. themes, wording and format of the items/tasks/questions) is consistent with the competencies to be assessed. Expert judgement is central.	Recognition purposes within the AQF (i.e. when assessing for summative purposes). This is crucial for ensuring industry credibility and acceptability, particularly when knowledge and understanding is required for competent performance in the workplace (e.g. where occupational health and safety issues are at stake).
Criterion related	Subdivides into concurrent and predictive validity: • Concurrent validity—concerned with comparability and consistency of a candidate's assessment outcomes with other related measures of competency • Predictive validity—concerned with the ability of the assessment outcomes to accurately predict the future performance of the candidate and how the candidate will be able to apply the knowledge and skills to new or other situations outside the context of the assessment event (i.e. transferability)	• Concurrent validity—evidence of competence on one task should be consistent with evidence of competence on another related task (e.g. on and off the job assessments that are measuring the same unit of competency should provide consistent evidence of competence levels). • Predictive validity—assessments should be able to predict if the candidate will be able to apply the relevant skills in knowledge in the workplace	Can be statistically determined (e.g. correlational analyses), or can be judged by the task developer through comparisons and follow-up studies with other measures.	• Concurrent validity is important for establishing transferability of the assessment outcomes. As such it is important to the National Training Board's (1992) original definition of competency. • Predictive validity is particularly important for employability contexts and selection processes.

Construct	Concerned with the theoretical evidence of what is being assessed. Constructs are non-observable qualities, such as attitudes and values, competencies and learning, which require inferences to be made by the assessor. A construct is a way of organising observations to help interpret them. *Construct validity is concerned with the* degree to which the evidence collected can be used to infer competence in the intended area, without being influenced by other non-related factors (such as literacy levels, etc.).	Observations of a driver's overall behaviour in heavy traffic, parking, hill starts, open road, night driving and at speed, allows us to infer that the candidate has a high ability in driving. Examples of where construct validity cannot be claimed include: • performance in a role-play *situation that is dependent upon* cultural and personality characteristics of the candidate that are unrelated to the competency that is intended to be assessed • a paper and pencil test designed to measure knowledge and understanding of OHS, which is also measuring literacy skills of the candidate, where literacy is not relevant to the competency of interest	Task developers investigate what qualities (i.e. knowledge and/or skills) an assessment task measures by determining the degree to which the intended constructs account for performance on the assessment. This can be empirically tested through statistical procedures to test the relationship between the assessment tasks and the intended competency to be measured (e.g. factor analyses). The task developers need to gather evidence across a range of contexts (as guided by the Range of Variable statements of the competency standards). They then need to demonstrate how the competency that is being assessed in each context is not affected by the context. Task developers need to examine possible errors in interpreting evidence including, but not limited to, the adequacy and appropriateness of the: • representation of the competency (i.e. whether the evidence collected has sufficient coverage of the competency of interest, including the Range of Variable Statements) • task format, administration and scoring procedures • language used	Construct validity in CBA is concerned with how well the evidence supports the claims about the competency being measured. Without construct validity, content and criterion validity are not possible.
Consequential	Concerned with the consequences of the use of the assessment information for all stakeholders (e.g. hidden agendas, funding influences, maintaining pre-established relationship and continued employability of candidate/assessor). These consequences may influence the way in which assessors make the judgements of competency.	The candidate provides a portfolio of work samples to be demonstrated for purposes of recognition of previous learning and skills. The assessor uses the evidence to make an inference of competence, and returns the portfolio. No further use is made of the material without the candidate's approval, otherwise consequential validity may be compromised.	The users of the assessment information make value judgement. They need to examine whether evidence about consequences is directly relevant to validity. The interpretation of evidence should not be influenced by the perceived consequences of the decision.	High accountability and stakes situations (e.g. budgetary considerations, promotion, etc.). This could be particularly pertinent to the VET sector if funding is dependent upon positive assessment outcomes.

Note: Although the task developers have the responsibility to provide content, construct and criterion validity evidence, it should be emphasised that the task users (i.e. assessors) have the ultimate responsibility for evaluating the quality of the validity evidence provided and its relevance to their own purpose, context and target group.

Sources: Athanasou 1997; Bennett 1993; Bernadin & Beatty 1984; Cropley 1995; Elliot 1994; Griffin & Gillis 1997; Hager Athanasou & Gonczi 1994; Howell, Begelo, Moore & Evory 1993; Linn 1993; Linn 1994; Linn, Baker & Dunbar 1991; Masters & McCurry 1990; Messick 1989 & 1992; Rudner 1994; Tanner 1997; Wilson et al. 1988.

Table 2: Reliability types—Definitions and relevance to CBA

Type of validity	Description	Examples	Determining validity	Importance to CBA
Inter-rater across assessors	Consistency of judgement across different assessors using the same assessment task and procedure.	Two independent assessors (for example, peer and workplace supervisor) make the same judgement of competency of the candidate using the same assessment task and procedure. That is, would another assessor reach the same conclusions?	Reliability can be determined through moderation and/or verification procedures (e.g. comparing judgements of two independent assessors who have not consulted each other on their decisions). It can also be statistically determined.	Extremely important in competency based assessments that have strong reliance on assessor judgements. Helps identify harsh and lenient assessors, clarity and consistency of interpretation of standards, assessment criteria, scoring procedures and decision making rules. Determined during the development stage of the assessment task, but continual review of variations across assessors may also be necessary.
Intra-rater (also referred to as test-retest) within assessors	Consistency of assessment outcomes across time and location, and using the same assessment task administered by the same assessor. Examines whether the assessment task leads to consistency of outcomes by the same assessor.	If the assessment was repeated on another day, using the same assessment task with the same candidate, would the same results be produced?	Reliability can be achieved through assessing the candidate(s) on more than one occasion using the same assessment task and context (e.g. location) during the pilot stage of the task development. The task developer would then need to compare assessment outcomes and decisions.	Important to determine when using multiple methods and multiple performances of candidates, all of which are highly encouraged in a CBA system. This form of reliability needs to be determined during the task development stage.
Parallel forms across tasks	Concerned with determining the equivalence of two alternative forms of a task.	On- and off-the-job assessments of the same unit of competency should produce consistent outcomes. Do the two assessment tasks produce consistent findings?	Administer two equivalent tasks to the same group of candidates and determine correlations among scores.	Important when the assessor has a selection of assessment for the same competency unit(s). This may be particularly important when an assessment system includes task bank facilities. Again, this form of reliability is pertinent to the task development and validation stage.
Internal consistency within task	Concerned with how well the items or tasks act together to elicit a consistent type of performance.	If using multiple assessment tasks to make an overall judgement of competence, one needs to examine whether the tasks are producing consistent evidence of competence levels. Are the sub-tasks acting in a consistent manner to make an overall judgement of competence?	Requires statistical procedures to be applied by the task developers to estimate reliability (Thorndike 1976). Often known as Cronbach Alpha, and applied to test or rating scale quantitative data.	Important for assessment of underpinning knowledge and understanding required for competent performance. This form of reliability is particularly important when there is a large range of items/tasks that can be selected by the assessor. Internal consistency need only be determined empirically during the development phase.

Sources: Athanasou 1997; Bennett 1993; Griffin & Nix 1991; Groth-Marnat 1990; Kerlinger 1973; Masters & McCurry 1990; Messick 1992; Rudner 1992; Rudner 1994; Wilson et al. 1988.

With the introduction of competency-based assessments that utilise performance assessments requiring judgements to be made by the assessor, 'inter-rater' and 'intra-rater' measures of reliability become increasingly pertinent (Bennett 1993). Table 2 provides a brief description of different types of reliability.

Unlike validity, it is not necessary to satisfy all types of reliability. Deciding which type of reliability to use will depend upon the nature of the competency to be assessed, the way it is assessed, and the purpose for which the assessment will be used. Determining reliability tends to be the responsibility of those developing and validating the assessment tasks, rather than of the assessor. It is also a property of the process and how error is controlled. When reporting reliability, it is important to report the evidence and the means of controlling extraneous influences.

Objective assessments

What are they?

OBJECTIVE TEST ITEMS refer to questions that require the candidate to select a response from a set of alternative responses constructed by the test developer (e.g. multiple choice, true/false and matching). The term 'objective' is used to describe such test items because it is thought that there is no judgement to be made by the assessor when scoring such items (Wilson et al. 1998). However, there is always an element of subjectivity involved in any assessment format. In objective tests, the subjectivity occurs in the selection and construction of items to be included in the test, as well as in the scoring procedures/answer keys established by the test developers (Messick 1992). In CBA, the popularity of checklists reflects a belief that ticks and crosses constitute an objective approach.

The extensive use of objective tests, particularly in standardised assessment and reporting systems, has a number of possible advantages, such as:

- ease of scoring
- cost-efficient scoring procedures
- ease of assessing a group of candidates at one time
- appearance of 'objectivity'—hence thought to reduce possible assessor bias
- standardised administration conditions that can be easily established and maintained
- appropriate application within both a normative and/or criterion referenced assessment system
- multiple ways to assess underpinning knowledge and understanding
- ease of determining validity through statistical procedures
- ease of determining estimates of reliability

Although there are a number of strengths associated with objective tests, both researchers and assessment practitioners have raised some concerns. For instance, Shannon (1991) and Wiggins (1991) argue that multiple choice tests can often lack *face validity*. The task of choosing the one 'best' answer on a multiple choice test may be very different from situations in the workplace for which there may not be a best alternative nor any known solution (Wiggins 1991). Multiple choice questions are, and should be, limited to assessing knowledge and understanding. Paper and pencil tests that are designed to predict workplace knowledge and understanding should be validated for appropriateness in the given context and purpose. The task developers must document these conditions.

Taylor (1993), Linn et al. (1991) and Messick (1992) argue that concerns about limitations of objective tests are a result of the content of the test items, rather than the format. The effectiveness of multiple choice tests to assess recall of facts as well as high order cognitive skills (e.g. understanding), is dependent upon the skills of both the assessment designers and those interpreting the results (Anastasi 1988). According to both Griffin and Nix (1991) and Messick (1992), objective test questions (particularly multiple choice items) are very difficult to write, and require skills in test construction and data analysis. Few teachers and trainers have had the necessary training to write valid and reliable test items. The importance of acquiring competencies in developing assessment tools has recently been recognised in the VET sector, with the introduction of specialised units in assessment that form part of the Diploma of Training and Assessment Systems (NAWTB 1998).

Implications for validity of objective assessments in competency-based assessment

The validity of any assessment will depend upon the purpose of the assessment and the way in which the evidence is interpreted and used by the key stakeholders. Table 3 (on page 18) illustrates how each type of validity applies to objective testing techniques, and how they might be enhanced.

Implications for reliability of objective assessments in competency-based assessment

The importance of ensuring the reliability of the assessment has largely been addressed through standardising the administration and scoring procedures, in an attempt to minimise or eliminate the influence of contextual influences and judgement on the assessment decision (i.e. minimise measurement error).

Hence, errors of measurement in objective tests are associated with the level of precision of the assessment task, the degree of standardisation of the administration and scoring procedures, and the internal characteristics of the candidate (e.g. fatigue from a long test).

Objective tests have established scoring procedures and decision-making rules for judgements. At a minimum, task developers should report evidence of the internal consistency of the test.

Table 3: Validity—Recommendations for design and validation of objective assessments

Type of validity	Recommendations for the design and validation of objective tests
Face	• Select or design questions that address knowledge and understanding needed in workplace situations.
Content	• Prepare and review detailed task specifications covering the knowledge and skills to be assessed (use content experts in the review). • Sample adequately (sufficiently) from the competency domain (i.e. skills or knowledge). • Determine whether the test, as a whole, represents the range of the skills and knowledge required for competent performance (refer to the range of variables and the evidence guides in the competency standards). • Include questions that assess beyond recall of facts. • If the competency also entails psychomotor skills, the tests should be used in conjunction with performance tasks such as simulations, role-plays, workplace activities, etc.
Criterion related—Concurrent and predictive	• Obtain empirical evidence of performance after the assessment event to establish predictive validity. • Document the link between the candidate's performance on the objective test with that of another task/test (i.e. workplace assessments) to establish concurrent validity; evidence must include a comparison of the candidate's performance on the test with performance on another criteria (e.g. other tests and teacher/supervisor ratings).
Construct	• Collect supporting evidence that the test is related to the specific competency intended to be assessed (through empirical testing). • Check whether unrelated factors are contributing to performance on the test (e.g. literacy skills, test-wiseness). • Compare test scores before and after implementation of training (after training expect higher scores). • Gather evidence across a range of contexts (refer to Range of Variable Statements); show how the competency is not affected by the context of the assessment. • Compare test results of two groups of individuals who are expected to perform differently on the test (e.g. a group from the industry concerned and another group from outside the industry).
Consequential	• Specify the purpose of the assessment in the test specifications. The type and use of the information needs to be agreed to by all declared stakeholders prior to any assessments (preferably in writing).

Sources: Athanasou 1997; Bennett 1993; Linn et al. 1991; Rudner 1994.

Performance assessments

What are they?

PERFORMANCE ASSESSMENTS INCLUDE a range of assessment methods, requiring candidates to perform a task(s), and/or create an answer or a product, to demonstrate their knowledge and skills (e.g. simulation, portfolio, role-play, and essays). They involve direct observation of candidate's behaviour and/or inspection of a product. They require performance of a specific activity or a constructed response, often over an extended period of time (Elliot 1994; Lam 1995; Linn 1994; Messick 1992). In the VET sector, performance assessments involve demonstration of competencies and/or learning outcomes, and can range from simple constructed responses (e.g. open-ended written and/or oral questions), to comprehensive demonstration or collections of work over time (e.g. a portfolio) (Elliot 1994). In general, these assessments are characterised by direct observation and judgement.

There are differing views on the definition and understanding of the term 'performance assessments'. Terms such as 'authentic' and 'portfolio' are often used as examples of performance assessments. For instance, portfolio refers specifically to the gathering of evidence produced across time (Paulson and Paulson 1991), whereas authentic refers to the nature of the assessment tasks and its match to the context (Elliot 1994). Authentic assessment methods promote face validity. They are particularly important in competency-based assessments because they look at the realistic nature of the assessment and how well the task resembles workplace activities (Howell et al. 1993; Shavelson 1994). In many instances, they may be a part of a workplace activity.

Within the Australian context of competency-based assessments, performance assessments are a way in which assessors can gather evidence of competence to make a judgement. Hayton and Wagner (1998) identified the following six attributes of performance assessments that are important within a CBA system:

- ❖ the assessment activity reflects the criterion activity or realistic workplace activity
- ❖ assessment is multi-dimensional, encompassing more than knowledge
- ❖ assessment can be a product or a process or both
- ❖ assessment spans a continuum from simple to complex activities
- ❖ assessment is open ended
- ❖ scoring requires human judgement (Hayton and Wagner 1998, p.71)

Performance assessments offer a number of possible advantages, including:

- ❖ greater face validity (due to 'authentic' nature of tasks)
- ❖ overcoming test wiseness associated with objective tests
- ❖ greater relevance and direct evidence of competence
- ❖ greater flexibility provided to assessors to contextualise assessment tasks
- ❖ increasing fairness because the tasks can be designed to cater for individual needs, especially minority groups and candidates with disabilities
- ❖ empowerment of the candidate in the assessment process (e.g. selection of methods, gathering evidence)
- ❖ opportunities for assessment of the process as well as the end product (Dais 1993; Lam 1995; Linn et al. 1991; Messick 1992; Wilson et al. 1988)

Despite the widespread belief among practitioners of these advantages, there are dangers involved in assuming that performance assessments measure higher order, cognitively complex competencies (Linn et al. 1991). Performance assessments are not necessarily fairer than objective tests, as they can introduce forms of bias associated with judgement error (Gillis et al. 1997; Linn et al. 1991). Price (1989) identified a number of judgement errors associated with performance assessments that influence the reliability of assessment outcomes. These included:

- ❖ *The halo effect*—influenced by characteristics/qualities of the assessee not related to the competencies of interest (e.g. appearance)
- ❖ *Assessor bias and inconsistencies*—assessor preferences/values influence the way in which information is interpreted and used
- ❖ *First impression or primacy error*—the tendency for an assessor to place a higher value on behaviour or performance that occurred early in an assessment period
- ❖ *The spillover effect*—assessors are influenced by past assessment outcomes, and give a similar assessment result for the current assessment regardless of the current evidence

- *Same as me or different from me*—assessors give higher ratings if a candidate has similar qualities or characteristics as the assessor
- *Central tendency*—when in doubt, assessors systematically judge candidates as average

The influence of judgement error associated with performance assessments therefore needs to be tested during the validation of any performance assessment task.

Implications for validity of performance assessments in competency-based assessment

Although performance assessments claim high face validity, this alone does not provide sufficient evidence of other types of validity. For instance, Tanner (1997) argues that despite the face validity, there is still a need to demonstrate that performing what is described as an authentic task indicates how well one will be able to apply such competencies outside the assessment event.

Lam (1995) raises concerns that performance assessments can create difficulties in making comparisons of assessment outcomes between individuals and tasks. Although comparisons amongst individuals are discouraged within a CBA system, the issue of comparability is particularly important when determining the concurrent validity of an assessment task. Concurrent validity is difficult to determine because of the complexities associated with developing highly contextualised assessment tasks of equal difficulty. The issue of predicting transferability of competencies outside the assessment event has also been raised as a concern, given the highly contextualised nature of the assessment tasks (Dais 1993; Howell et al.1993; Lam 1995; Tanner 1997; Taylor 1993; Messick 1992). There is a dilemma: the more the tasks are contextualised the less it is possible to generalise about the transferability of the competencies.

However, Athanasou (1997) argues that evidence of concurrent validity is easier to collect than that of predictive validity in the VET sector, as the latter requires follow-up studies of trainees. He suggests that concurrent validity can be determined through examining performance of students on other subjects in a training course, or making comparisons with work placements or on-the-job training that may be part of the course. Given that many VET programs include work placements, criterion-related validity evidence should, in many instances, be able to be gathered cost-effectively.

Table 4 provides guidelines for designing and validating performance assessments.

Table 4: Validity—Recommendations for design and validation of performance assessments

Type of validity	Recommendations for the design and validation of objective tests
Face	• Ensure that tasks are guided by workplace rules, norms, expectation and restrictions to ensure workplace acceptance and confidence in the assessment activities. • Use direct observation of workplace activity. • Report to the stakeholders the clear link between the competency to be assessed and the required evidence to be collected when using indirect forms of evidence such as portfolio and simulation methods.
Content	• Use multiple tasks and multiple sources of evidence as the basis for competency judgements. • Develop task specifications that ensure all components of competency are addressed. • Involve experts in both the design and review of tasks.
Criterion related—Concurrent and predictive	• Gather evidence of post-assessment performance through follow-up studies with supervisors, the candidate and/or other assessors, to determine how accurately the assessment predicted the candidate's ability to apply the competencies to workplace settings. • Use a range of similar assessment tasks that have demonstrated equal complexity. • Adopt an integrated approach to assessment (that assesses all components of competency), as opposed to an atomistic, checklist approach that considers each performance criterion as a task within itself. • Use external assessors to provide independent assessments (using similar tasks).
Construct	• Examine the relationship between different sources of evidence of components of competency (perform, manage, transfer, handle contingencies and job/role environment skills): similarities in outcomes would indicate high levels of construct validity. • Compare assessment outcomes before and after implementation of training. • Compare assessment outcomes of two known groups who should perform differently on the assessment task (e.g. specialists versus hobbyists). • Gather evidence across a range of contexts (refer to Range of Variable Statements) and demonstrate how the competency is not affected by the context of the assessment.
Consequential	• Promote transparency of the assessment process through clear documentation and communication of: – the purpose of the assessment – the evidence to be collected – the way in which the evidence will be interpreted – how and what information will be reported to stakeholders

Sources: Bennett 1993; Linn et al.1991.

Implications for reliability of performance assessments in competency-based assessment

When establishing reliability of performance assessment, it is necessary to identify the likely sources of error. Although performance assessments are subjected to the same sources of error as that of objective tests, there are additional sources of measurement error that need to be considered. These include the influence of the judge, the context of the assessment, and the range and complexity of the task(s). All can influence the assessment judgement, and may distract the assessor from measuring the 'true' competence of the candidate. Therefore, establishing reliability of

performance assessment creates challenges for both task developers and assessors.

Factors that have been reported to influence the reliability of performance judgements include the:

- assessor's relationship with the candidate (Kingstrom and Mainstone 1985)
- characteristics of the candidate (Dobbins, Cardy and Truxillo 1988; Hollenbeck, Illgen, Phillips and Hedlund 1994; Oppler, Campbell, Pulakos and Borman 1992; Ritts, Patterson and Tubbs 1992)
- candidate's past performance (Murphy, Blazer, Lockhard and Eisenman 1985)
- assessors attitudes toward, and pre-conceived expectations of, the candidate (Diboye 1985)
- motivational factors such as hidden agendas, avoidance of conflict, budgetary factors and/or friendships (Hauenstein 1992; Robbins and DeNisi 1994)

In a recent study of factors influencing judgements in competency-based assessments, Gillis et al. (1997) found that greater exposure to industry audits, external verifications and accountability were associated with increases in the consistency of judgements. In low accountability conditions, assessors were influenced by non-performance-related characteristics of the candidate, such as physical appearance, age and length of time in the company.

There was also a misunderstanding of the notion of fairness amongst assessors. Some assessors adjusted the required level of performance of individuals with special needs, as opposed to altering the methods of gathering evidence. Lam (1995) suggests that to ensure fairness of assessment, assessors need to design individualised performance assessments that address the purpose, context, background and the individual needs of the candidate, without altering the performance levels required. There is a need for verification of assessment judgements to minimise the presence of assessor bias and any relaxation of standards. Such quality assurance procedures should be a central feature of all CBA systems, and made widely known to, and applied by, all assessors.

Given that professional judgement of the assessor is a major attribute of performance assessments within a CBA system (Hayton and Wagner 1998), task developers should ensure that evidence of inter-rater reliability is established and documented. Particular attention needs to be given to the clarity of the evidence to be gathered and the decision-making rules, so that consistency of judgements can be facilitated.

Synthesis of information

A NUMBER OF LESSONS can be learnt from studies that have explored issues of reliability and validity of assessments. This paper has identified appropriate types of reliability and validity relevant to both performance assessments and objective tests that can be used within a CBA system.

Validity

Within the VET sector, performance assessment tasks tend to have greater face validity than objective tests. This is due to the tendency for performance assessments to rely on direct observation and the authentic nature of the task. Even though it is more difficult to establish face validity using paper and pencil tests to assess skills, its importance rests with assessment of underpinning knowledge and understanding. Face validity is essential. Without it, assessments lack credibility.

Content validity is a familiar concept, and incorporates common assessment practices. It is both relatively easy and important for the task developers to establish adequate content coverage of an objective test to measure the underpinning knowledge and understanding reflected in the industry standards. However, it may be harder to achieve with performance assessments, as they are often developed for specific contexts, purposes and individual characteristics of candidates. Assessors need to sample a sufficient and adequate range of evidence of competent performance.

Criterion validity tends to be broken down further into two major categories: concurrent and predictive. Concurrent validity is important for examining transferability of competence to new or related contexts as evidence is gathered from multiple sources, whether that is individual objective test items or performance tasks. However, difficulties can emerge when making comparisons across varying tasks/items that have been too highly contextualised. The assessor needs to gather evidence regarding the degree to

which the skills and knowledge transfer to other tasks/items. Predictive validity is thought to be simple and straightforward, and important for any type of assessment—but follow-up studies are often under-resourced, particularly with localised assessments. It is, however, extremely important in competency-based assessments, since every assessment must predict application and transferability of competencies to the workplace. Without an emphasis on continued demonstration of competence in the workplace, CBA is of questionable value.

Construct validity is often thought to be the most difficult to both understand and achieve. It provides a framework within which to interpret the evidence gathered from the assessment tasks. Its importance rests with assessment of competencies that cannot be directly observed—that is, when the task is not assumed to be a direct measure of the competency or that they are one and the same thing. For example, if assessing all components of competency (performing the task skill, managing a number of tasks, job/role environment skills, and contingency management skills), including transferability to new contexts, construct validity becomes particularly pertinent.

Finally, consequential validity is particularly important in high stakes assessments (e.g. recruitment or selection purposes). For instance, trainers/assessors may teach to the test. This can lead to inadequate coverage of the competencies, which impacts on content and construct validity. In performance assessments, for instance, consequential validity may be compromised given the high subjectivity of the judgement involved and the procedures used (particularly if non-standardised). The assessor may distort the judgement, decision and recommendations.

Validity is not simply a property of the assessment task in isolation. An assessment task that is highly valid for one use or context may be invalid for another.

Reliability

It is not necessary to satisfy all types of reliability. The decision as to which type of reliability evidence to use will depend upon both the nature of the competency to be assessed, and the purpose for which the assessment will be used. The responsibility for determining reliability tends to rest with the task developers. Typically, traditional types of reliability that are concerned with the task (e.g. internal consistency) have been the focus of attention with objective tests. With performance tasks, it is important to establish both inter- and intra-rater reliability. These two types of reliability are concerned with the consistency of assessors, as opposed to consistency of the tasks (i.e. parallel forms and internal consistency).

Inter-rater reliability is extremely important in tasks that rely on assessor judgements. This type of reliability is concerned with the consistency *across assessors*, and focuses upon:

- collection of evidence
- interpretation and synthesis of the evidence
- agreement among resultant judgements

Evidence of inter-rater reliability can help identify harsh and lenient assessors, and variations of understanding and application of standards, assessment criteria, scoring procedures and decision-making rules. Decision-making rules and scoring procedures need to be pre-determined with all forms of assessment. This will bolster inter-rater reliability. However, reliance on inter-rater reliability as the sole indicator of good quality assessments may be flawed, as it may be possible that all assessors consistently interpret and judge competence and the standards inaccurately.

When conducting assessments for the purpose of prediction, Groth-Marnat (1990) suggests that the intra-rater technique is the most preferred, since it gives an estimate of the dependability of the assessment from one administration to the next. It is concerned with consistency *within the assessor* (i.e. will the assessor make the same judgement using the same task and context at a different time?). Again, caution needs to be exercised as the same assessor could be consistently misinterpreting the evidence. Intra-rater reliability should only be used when the competency being measured is relatively stable across time, and when the interval between the assessments is short. The task developers need to allow for assessor and candidate experience, practice effects, and learning that may have taken place in between the two assessments, when considering intra-rater reliability.

Parallel form reliability estimates are concerned with the consistency of evidence produced from alternative assessment tasks (i.e. *across* tasks). In CBA, parallel forms can readily be used where the assessor can sample from a range of equivalent tasks to assess against a unit of competency. There are a number of limitations to this estimate of reliability. It may place extra demands on the candidates and assessor during the development and validation phase. For instance, during the trialling of the assessment tasks, the candidate(s) would need to complete a number of tasks/tests.

Internal consistency estimates of reliability are concerned with the accuracy and consistency of evidence collected *within* a task. In theory it works like this: If a task is randomly split into sub-tasks, will the sub-tasks provide a common outcome? If these two sub-tasks are again split, how consistent are the outcomes? If these are split again, how consistent are these numerous outcomes? Internal consistency is a measure of the proportion of consistent

outcomes. Internal consistency estimates are important if the assessor uses a set of tasks to make an overall judgement of competence. The assessor needs to examine whether the tasks are producing consistent evidence of competence. This type of reliability is most commonly determined through statistical analysis, but task developers could apply the procedures outlined above to estimate internal consistency of performance assessments.

Given that competency-based assessments require judgements to be made by assessors using both objective tests and performance assessments, it may be better in a CBA context to discuss reliability in terms of how judgement errors and contextual influences can be controlled and accuracy improved thereby.

Implications for competency-based assessment

The issue is not which form of assessment may be more appropriate for use within a CBA system. Rather, it is the appropriateness and importance of the different types of reliability and validity which need to be evaluated, according to the purposes of the assessment and the way in which the evidence will be interpreted and used by the assessor and the stakeholders.

Central to any assessment and reporting process is the evidence. The technical criteria for evaluating the assessment must correspond to each stage within the assessment and reporting model (i.e. the purpose, the evidence collected, the way in which judgements are made, and what the information will be used for), regardless of the nature or form of the assessment. The purpose of the assessment, and the stakeholders' reporting requirements, will define the type of evidence that needs to be collected, which, in turn, influences how assessors use and interpret the information to make a judgement.

Evidence is also crucial in reliability and validity of assessments. The methods used to collect the evidence will impact on the reliability, whilst the way in which assessors use and interpret the evidence collected will impact on validity. As reliability creates the foundation for validity, an assessment should aim to reduce the error or 'noise' in the evidence collected or used.

Sources of error in objective tests are associated with the:

- method of gathering evidence (i.e. the level of precision of the assessment task and the degree of standardisation of the administration and scoring procedures)
- characteristics of the candidate (e.g. fatigue from a long assessment)

In performance assessments, there are additional sources of error:

- ❖ characteristics of the assessor (e.g. preconceived expectations of the competency level of the candidate)
- ❖ context of the assessment (e.g. location)
- ❖ range and complexity of the task(s) (e.g. the level of contextualisation of the task)

Each of these factors need to be controlled throughout the assessments, to improve the reliability. Reliability can be increased by controlling the way in which the evidence is collected (e.g. standardising the administration and scoring procedures).

When establishing reliability and validity, not all types are important at all assessment and reporting stages (Griffin and Nix 1991). Knowledge of when these different types come into play will help the assessor to incorporate procedures to both establish and enhance reliability and validity. This is illustrated in table 5.

The table illustrates that when establishing validity, the process of gathering and interpreting evidence appears to be the most crucial component of the assessment and reporting process. Similarly, the interpretation of evidence, as well as the way in which it was gathered, influences reliability. Assessors need to take into account the way in which evidence is collected, interpreted, synthesised and evaluated, to make an overall valid and reliable judgement of competence. In simple terms, validity is associated with the use and interpretation of the evidence collected, whilst reliability is concerned with precision and accuracy of the evidence and procedures used.

Table 5: Linking aspects of CBA processes to reliability and validity

	Validity						Reliability			
Assessment and reporting process	**Face**	**Content**	**Construct**	**Predictive**	**Concurrent**	**Consequential**	**Inter-rater**	**Intra-rater**	**Parallel forms**	**Internal consistency**
Purpose				✔		✔				
Evidence	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔
Methods	✔	✔					✔	✔	✔	✔
Interpretation		✔	✔	✔	✔	✔	✔	✔		
Judgement						✔	✔	✔		
Recording and reporting	✔			✔		✔				

Note: Reliability is linked to neither the purpose nor reporting stages of the assessment process.

Guidelines for establishing procedures to enhance reliability and validity

THE REVIEW OF the literature suggests a number of ways in which reliability and validity can be enhanced within competency-based assessments.

Validity

Validity of an assessment refers to use and interpretation of the evidence collected, as opposed to the assessment method or task per se. Hence, validity is not simply a property of the assessment task.

Face validity

- ❖ select and/or design assessment tasks that are based on or resemble workplace contexts and situations
- ❖ if using indirect forms of evidence, such as portfolios and simulations, report to the stakeholders the clear link between the competency to be assessed and the required evidence
- ❖ include the stakeholders in the selection of appropriate methods

Content validity

- ❖ determine whether the assessment task as a whole represents the full range of the knowledge and skill specified within the unit(s) of competency (refer to the range of variables and the evidence guides in the competency standards)
- ❖ prepare and review detailed task specifications covering the knowledge and skills to be assessed
- ❖ involve content experts in both the assessment task design and the review of the tasks' match to the competency(ies)
- ❖ use more than one task and source of evidence as the basis for judgment

Construct validity

- collect supporting evidence (empirical and theoretical) that the assessment task is related to the specific competency intended to be assessed
- examine different sources of evidence of knowledge and skills that are thought to underpin the competency
- compare assessment outcomes before and after learning: evidence of improvement of the candidate's performance on the assessment task would provide an indication of construct validity, when compared to a group that has not had the opportunity to acquire the relevant skills and/or knowledge
- compare assessment outcomes of two groups who are known to perform differently on an assessment task: select one group that is considered to be proficient in the competency, and compare their performance to those of the lower expecting group (e.g. specialist versus hobbyist)
- adopt an integrated approach to assessment, as opposed to the atomistic approach
- gather evidence across a range of contexts (refer to the Range of Variable Statements of the competency unit[s]); show how the competency assessed is not affected by the context

Criterion validity

Predictive

- during the validation of the assessment tasks gather evidence of performance at a later date, through follow-up studies/interviews with supervisors, the candidate and/or other assessors
- collect evidence of a variety of performances over time
- collect evidence of transferability of competence to new contexts and situations

Concurrent

- compare the candidate's performance of the assessment task with another measure of competency, such as self-assessment or on-the-job assessment
- compare assessment outcomes from a range of similar tasks that have demonstrated equal complexity
- use external assessors to provide independent assessments (using similar tasks)

Consequential validity

- identify the purpose, boundaries and limitations of the interpretations that can be made of evidence collected
- establish clear documentation and communication for all stages of the assessment and reporting process, including:
  - the purpose of the assessment
  - the evidence to be collected
  - the way in which the evidence will be interpreted and judged
  - how and what information will be reported to and used by stakeholders

Reliability

Reducing the amount of error within the evidence will increase reliability. The following guidelines will assist in making a correct judgement of the candidate's competence. In general, standardising the type of evidence to be gathered, how it will be gathered, and how it will be interpreted, will increase reliability.

Inter-rater reliability (across assessors)

- moderate assessment judgements with internal or external assessors
- maintain a representative sample of assessment tasks to compare from context to context/year to year
- use a panel of independent assessors to evaluate the sample of assessment tasks
- establish and document clear assessment procedures/instructions for collecting, analysing and recording outcomes to evaluate the evidence collected, the circumstances under which it was collected, and the extent to which the procedures were followed
- use multiple tasks and multiple sources of evidence as the basis for judgement
- develop exemplar assessment tasks and procedures as models for assessors
- document the qualifications and experience required of assessors, and describe any training in assessment that needs to be undertaken

Intra-rater reliability (within assessor)

- inform assessors about common sources of judgement error in CBA, and encourage self-awareness of own biases
- minimise the time between two assessments, to avoid other factors (i.e. learning) that may occur between the two occasions

Parallel forms reliability (across tasks)

- ensure the same level of difficulty of two assessment tasks that have been developed for the same target group and competency unit/s
- use verifiers to conduct independent assessments using similar tasks

Internal consistency reliability (within tasks)

- break the task into sub-components, and check agreement among these sub-components
- increase the number of assessment tasks used to make the decision
- design and use assessment tasks that have a range of difficulty levels

Validation of an assessment process should integrate the various forms of reliability and validity evidence outlined above, and will require the assessment task developers and users (i.e. assessors) to make an holistic judgement as to whether the reliability and validity evidence supports the intended use and interpretation of assessment evidence for the specified purpose(s). These also need to be reported to potential users.

Ultimately, the validation of an assessment in terms of reliability and validity requires evidence of careful task development, clear and concise assessment criteria against the competency standards, appropriate task administration procedures, and adequate scoring/decision-making rules and recording procedures. A clear distinction is needed between validation and endorsement.

Findings and directions for further research

THIS REVIEW HAS revealed a number of areas requiring further research investigation. These can be summarised under the following four headings.

Validation approaches used by workplace assessors and VET practitioners within Australia. The majority of studies on reliability and validity reported in this review have been based on large-scale testing programs within the United States of America. Research needs to document the way in which assessments are currently being validated (in terms of reliability and validity) in Australia—not only for large-scale purposes (such as exemplar tasks to be included in the non-endorsed components of industry training packages and/or licensing arrangements), but also for local purposes.

Transferability of competencies outside the assessment event. In particular, can highly contextualised assessment tasks adequately predict competent performance in the workplace across a range of settings and contexts? Despite the fact that portability and recognition of qualifications is dependent upon the ability to transfer the competencies to new situations and contexts—all of which underpin the success of the ARF—there is very little empirical research that has investigated this concept.

Consequences of competency-based assessments in both vocational educational settings and the workplace. The research should identify both the intended and unintended effects of competency-based assessments upon the way in which training is delivered, and how the assessment process is carried out, judgements made and the outcomes reported.

Factors that influence judgements in CBA and how such factors impact on reliability and validity. With the large uptake of performance assessments in the VET sector, research should be conducted to identify how judgements of competency are made, what factors influence such judgements, and how these factors impact on reliability and validity.

References

Athanasou, JA 1997, *Introduction to educational testing*, Social Science Press, NSW.

Anastasi, A 1988, *Psychological testing*, 6th edition, Macmillan, New York.

Bennett, Y 1993, 'The reliability and validity of assessments and self-assessments of work-based learning', *Assessment and evaluation in higher education*, vol.18, no.2, pp.3–16.

Bernadin, HJ & Beatty, RW 1984, *Performance appraisal: Assessing human behaviour at work*, Kent Publishing, Boston.

Bloch, B, Clayton, B & Favero, J 1995, 'Who assesses?' in *Key aspects of competency-based assessments*, ed WC Hall, NCVER, Adelaide.

Bowers, BC 1989, 'Alternatives to standardized educational assessment', *ERIC digest series*, ERIC Clearinghouse on Educational Management, Oregon, USA.

Clayton, B 1995, *Focussing on assessment: Strategies for off-job teachers and träiners*, NCVER, Adelaide.

Cronbach, LJ 1971, 'Test validation', in *Educational measurement*, ed. RL Thorndike, 2nd edition, American Council on Education, Washington DC, USA, pp.443–507.

—— 1984, *Essentials of psychological testing*, 4th edition, Harper & Row, New York.

Cropley, M 1995, 'Validity', in *Key aspects of competency-based assessment*, ed. WC Hall, NCVER, Adelaide, in association with the DEET.

Dais, TA 1993, 'An analysis of transition assessment practices: Do they recognise cultural differences?', *Selected readings in transition: Cultural differences, chronic illness, and job matching*, vol.2, Illinois, USA.

Diboye, RL 1985, 'Some neglected variables in research on discrimination in appraisals', *Academy of management review*, vol.10, pp.166–127.

Dobbins, GH, Cardy, RL & Tuxillo, DM 1988, 'The effects of purpose of appraisal and individual differences in stereotypes of women on sex differences in performance ratings: A laboratory and field study', *Journal of Applied Psychology*, vol.73, no.3, pp.551–558.

Elliott, SN 1994, 'Creating meaningful performance assessments: Fundamental concepts', the Council for Exceptional Children, Vancouver, ERIC Clearinghouse on Disabilities and Gifted Education, Vancouver, ERIC/OSEP Special Project on Integrating Information Dissemination.

Gillis, S, Griffin, P, Trembath, R & Ling, P 1997, 'The examination of the theoretical underpinning of assessment', A report of a research funded by the Australian National Training Authority Research Advisory Council, unpublished, University of Melbourne, Melbourne.

Gillis, S, Keating, J & Griffin, P 1998. *Best practice in assessment in school industry programs. Draft research report: Stages 1 & 2,* The Australian Student Traineeship Foundation, Sydney (in press).
Glaser, R 1981, 'The future of testing: A research agenda for cognitive psychology and psychopathics', *American Psychologists,* vol.36, no.9, pp.9–23.
Gonczi, A, Hager, P & Athanasou, JA 1993, *The development of competency-based assessment strategies for the professions,* National Office of Overseas Skills Recognition, Research paper no.8, June, AGPS, Canberra, in association with the DEET.
Griffin, P 1995, 'Competency assessment: Avoiding the pitfalls of the past', *Australian and New Zealand Journal of Vocational Education,* vol.3, no.2, pp.33–59.
—— 1997, 'Developing assessment in schools and workplace', Paper presented at the Inaugural Professorial Lecture, Dean's Lecture Series, Faculty of Education, University of Melbourne, Melbourne, September 18.
—— 1998, *Measuring achievement using sub-tests from a common item pool: An application of the Rasch Model,* IIEP Paris (in press).
Griffin, P & Gillis, S 1997, *Workplace assessor training manual,* University of Melbourne, Melbourne.
Griffin, P & Nix, P 1991, *Educational assessment and reporting: A new approach,* Harcourt Brace Jovanovich, NSW.
Groth-Marnat, G 1990, *Handbook of psychological assessment,* 2nd edition, John Wiley & Sons, Canada.
Guthrie, H 1993, 'Assessing competence and appraising individuals: Linking the concepts through competency-based training', Paper presented at Testing Times, NCVER Conference, Adelaide.
Hager, P, Athanasou, J & Gonczi, A 1994, *Assessment: Technical manual,* AGPS, Canberra, in association with the DEET.
Hager, P, Gonczi, A & Athanasou, J 1994, 'General issues about assessment of competence', *Assessment and Evaluation in Higher Education,* vol.19, no.1, pp.3–16.
Hayton, G & Wagner, Z 1998, 'Performance assessment in vocational education and training,' *Australian and New Zealand Journal of Vocational Education Research,* 6(1), 69-85.
Hauenstein, NM 1992, 'An information-processing approach to leniency in performance judgements', *Journal of Applied Psychology,* vol.77, no.4, pp.485–493.
Hollenbeck, JR, Illgen, DR, Phillips, JM & Hedlund, J 1994, 'Decision risk in dynamic two stage contexts: Beyond the status quo', *Journal of Applied Psychology,* vol.79, no.4, pp.592–598.
Howell, K, Bigelow, S, Moore, E, & Evoy, A 1993, 'Bias in authentic assessment', *Diagnostique,* vol.19, no.1 pp.387–400.
Kerlinger, FN 1973, *Foundations of behavioural research,* 3rd edition, Holt Rinehart and Winston, New York.
Kingstrom, PO & Mainstone, LE 1985, 'An investigation of the rater-ratee acquaintance and rater bias', *Academy of Management Journal,* vol.28, pp.641–653.
Lam, TCM 1995, 'Fairness in performance assessment', *Eric digest,* ERIC Clearinghouse on Counselling and Student Services, NC, USA.

Linn, RL 1993, 'Educational assessment: Expanded expectations and challenges', *CSE technical report 351*, National Centre for Research on Evaluation, University of California, Los Angeles.
—— 1994, 'Evaluating the technical quality of proposed national examination systems', *American Journal of Education*, vol.102, August, pp.565–579.
Linn, RL, Baker EL & Dunbar, SB 1991, 'Complex, performance-based assessment: Expectations and validation criteria', *Educational Researcher*, vol.20, no.8, November, pp.15–21.
Masters, GN & McCurry, D 1990, *Competency-based assessment in the professions*, AGPS, Canberra.
Messick, S 1989, 'Validity', in *Educational measurement*, ed. RL Linn, 3rd edition, American Council on Education, Macmillan, New York.
—— 1992, 'The interplay of evidence and consequences in the validation of performance assessments: Research report', Paper presented to the Annual Meeting of the National Council on Measurement in Education, San Francisco.
Murphy, KR, Balzer, WD, Lockhard, MC & Eisenman, EJ 1985, 'Effects of previous performance on evaluations of present performance', *Journal of Applied Psychology*, vol.70, pp.72–84.
National Assessors and Workplace Trainers Body 1998, *The training package for assessment and workplace training*, ANTA, Melbourne.
NTB (National Training Board) 1992, *Policy and guidelines*, 2nd edition, NTB, Canberra.
Oppler, SH, Campbell, JP, Pulakos, ED & Borman, WC 1992, 'Three approaches to the investigation of subgroup bias in performance measurement: Review, results and conclusion', *Journal of Applied Psychology*, vol.77, no.2, pp.201–217.
Paulson, FL & Paulson, PR 1991, 'The ins and outs of using portfolios to assess performance: Revised', Expanded version of a paper presented at the Joint Annual Meeting of the National Council of Measurement in Education and the National Association of Test Directors, Chicago, Illinois, USA.
Price, M 1989, 'Improving personnel evaluation process: A synthesis of research and practice', Paper presented at the Annual Meeting of the Southern Regional Council on Educational Administration, Columbia, SC, November.
Ritts, V, Patterson, M & Tubbs, ME 1992, 'Expectations, impressions and judgements of physically attractive students: A review', *A Review of Educational Research*, vol.62, no.4, pp.413–426.
Robbins, TL & DeNisi 1994, 'A closer look at interpersonal affect as a distinct influence on cognitive processing in performance evaluations', *Journal of Applied Psychology*, vol.79, no.3, pp.341–353.
Rudner, LM 1992, 'Reducing errors due to the use of judges', *ERIC/TM digest*, American Institute for Research, Washington DC, USA.
—— 1994, 'Questions to ask when evaluating tests', *ERIC/AE digest*, ERIC Clearinghouse on Assessment and Evaluation, Washington DC, USA.
Shannon, DM 1991, 'Teacher evaluation: A functional approach', Paper presented at the Annual General Meeting of the Eastern Educational research Association, 14th Boston, Massachusetts, USA.
Shavelson, R 1994, 'Performance assessment', *International Journal of Educational Research*, vol.21, no.3, pp.235–244.

Tanner, DE 1997, 'The long (suit) and the short (comings) of authentic assessment', Paper presented at the Annual Meeting of the American Association of Colleges for Teacher Education, Phoenix, Arizona, USA.
Taylor, D 1993, 'Reassessing performance based assessment', Paper presented at the Annual Conference of the Missouri Unit of the Association of Teacher Education, Osago Beach, Montana, USA.
Thorndike, RL 1976, *Educational measurement*, American Council of Education, Washington DC, USA.
—— 1988, 'Reliability', in *Educational research methodology and measurement: An international handbook*, ed. JP Keeves, Pergamon Press, Oxford, England, pp.330–343.
Wilson, LR, Scherbarth, BC, Brickell, HM, Mayo, ST, & Paul, RH 1988, 'Determining reliability and validity of locally developed assessments', ERIC Clearinghouse, Illinois, USA.
Wiggins, G 1991, 'A response to Cizek', *Phi Delta Kappan*, vol.72, no.9, pp.700–703.
Woehr, D & Roch, S 1996, 'Context effects in performance evaluation: The impact of ratee sex and performance level on performance ratings and behavioural recall', *Organisational Behaviour and Human Decision Processes*, vol.66, no.1, pp.31–44.
Zeller, RA 1988, 'Validity', in *Educational researcher, methodology, and measurement: An international handbook*, ed. JP Keeves, Pergamon Press, Oxford, England, pp.322–329.

Other titles in the series **Review of Research**

A brief history of the evaluation of VET in Australia, R McDonald, G Hayton

Alternative VET pathways to indigenous development, R Boughton

Assessor training programs, R Docking

The changing nature and patterns of work and implications for VET, P Waterhouse, B Wilson, P Ewer

Competition and market reform in the Australian VET sector, D Anderson

Developing the training market of the future, NCVER

Entry-level training, D Lundberg

Flexible delivery of training, P Kearns

Globalisation and its impact on VET, B Hobart

The impact of generic competencies on workplace performance, J Moy

Impediments to the employment of young people, M Wooden

The internationalisation of vocational education and training, P Smith, S Smith

Learning in the workplace, P Hager

Public and private training provision, K Barnett

Quality assurance in VET, P Hager

Returns to enterprises from investment in VET, S Billett, M Cooper

Vocational education and training for people from non-English-speaking backgrounds, V Volkoff, B Golding

Vocational education and training for people with disabilities, N Buys, E Kendall, J Ramsden

Vocational education and training in rural and remote Australia, S Kilpatrick, R Bell

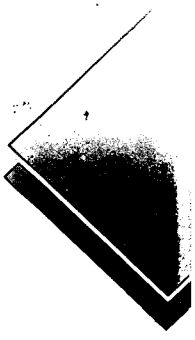
Vocational education and training in Australian correctional institutions, B Semmens, J Oldfield

Vocational education in schools, R Ryan

This review of research is one of a series of reports commissioned to draw conclusions from the research on key topics in vocational education and training.

Shelley Gillis is a research officer at the Assessment Research Centre, University of Melbourne. Her previous research includes the development of the endorsed components of the Training Package for Assessment and Workplace Training, examining factors influencing judgements in competency-based assessment and identifying best-practice models for assessment of school-industry programs.

Andrea Bateman is the foundation manager of the Ballarat Assessment Centre, University of Ballarat. She is also a training recognition consultant and auditor for OTFE. Her previous research includes the development of Training Packages for CREATE, Rural Training Council of Australia and National Assessors and Workplace Trainers Body.



ISBN 0-87397-542-1

9 780873 975421