Running head: ASSESSMENT QUALITY

This article was published as

Baartman, L. K. J., Bastiaens, T. J., Kirschner, P. A., & Van der Vleuten, C. P. M. (2007). Evaluation assessment quality in competence-based education: A qualitative comparison of two frameworks. ***Educational Research Review, 2,*** **114-129.**

Evaluating Assessment Quality in Competence-Based Education: A Qualitative Comparison of Two Frameworks

Liesbeth K.J. Baartman[ab]*, Theo J. Bastiaens[bc], Paul A. Kirschner[ab], Cees P.M. van der Vleuten[d]

[a]Utrecht University, the Netherlands

[b]Open University of the Netherlands

[c]Fernuniversität Hagen, Germany

[d]Maastricht University, the Netherlands

This research was supported by the Netherlands Organisation for Scientific Research (NWO) under project number PROO 411-02-363

* Correspondence concerning this article should be addressed to: Liesbeth Baartman, Utrecht University, Department of Educational Sciences, P.O. Box 80140, 3508 TC, Utrecht, The Netherlands. E-mail: L.K.J.Baartman@uu.nl

Abstract

Because learning and instruction are increasingly competence-based, the call for assessment methods to adequately determine competences is growing. Using just one single assessment method is not sufficient to determine competence acquisition. This article argues for Competences Assessment Programmes (CAPs), consisting of a combination of different assessment methods, including both traditional and new forms of assessment. To develop and evaluate CAPs, criteria to determine their quality are needed. Just as CAPs are combinations of old and new forms of assessment, criteria used to evaluate CAP quality should be derived from both psychometrics and edumetrics. A framework of ten quality criteria for CAPs is presented, which is then compared to Messick's framework of construct validity. Results show that the 10-criterion framework partly overlaps with Messick's, but adds some important new criteria, which get a more prominent place in quality control issues in competence-based education.

Keywords: evaluation criteria; quality control; assessment programmes; competence-based education

Evaluating Assessment Quality in Competence-Based Education: A Qualitative Comparison of Two Frameworks

Modern societies have dramatically changed due to technological changes such as the development of information technology systems. Service industries have become knowledge oriented, production economies have become knowledge economies and production workers have become knowledge workers. Learners need to be flexible and adaptive if they are to function well in today's complex and global societies. To support the needs of these new learners, education is changing its focus from one of transmitting isolated knowledge and skills to one of acquiring complex competences, guiding learners in developing skills for learning and getting information from the diverse range of sources available in modern society. In short, education is increasingly becoming learner-centred and competence-based.

As part of the larger drive to change the curriculum, assessment needs to be reformed as well. Biggs' (1996) idea of constructive alignment between instruction, learning and assessment implies that these three elements should be based on the same underlying principles, in this case competence-based education. Birenbaum et al. state in their EARLI position paper (2006) that current assessment practices in European countries fail to address learners' needs because they tend to focus on assessment *of* learning instead of on assessment *for* learning, are limited in scope, drive teaching for *assessment* instead of teaching for *learning*, and ignore individual differences. Although part of this might be true, new assessment methods are not without problems either and some feel that the evidence against classical tests is not as strong as has been claimed (Hambleton & Murphy, 1992), and that the claim that newer forms of assessment are better suitable to address learners' needs still needs empirical confirmation (Stokking, Van der Schaaf, Jaspers, & Erkens, 2004). Still, as a consequence of the changes towards competence-based education, a call is growing for the development of assessment methods that can adequately determine competence acquisition. The innovation of assessment might even be the cornerstone of success for the implementation of competence-based education (Tillema, Kessels, & Meijers, 2000). Studies have shown that no greater impulse for learning exists than assessment

(Frederiksen, 1984) and that a strong relationship exists between learning and assessment, implying that what is assessed strongly influences what is learned (e.g., Alderson & Wall, 1993). In other words, if European countries want to reform their curricula, assessment must have an important place in the reform process and assessment approaches need to focus on the integrated assessment of knowledge, skills and attitudes.

Though it is clear that assessment needs re-thinking in the direction of competence assessment, it is not clear what requirements should be used for these new competence assessments. This is an important question to address, as the quality of assessment is increasingly being regarded as a very important element of the quality of education as a whole. Assessments in competence-based education may require new and other quality criteria to evaluate them. These criteria need to be more compatible with the principles and ideas of competence-based education. The goal of this article is to provide a first step towards the solution of this problem. A framework of ten quality criteria for competence assessment (Baartman, Bastiaens, Kirscher, & Van der Vleuten, 2006; in press) is compared to Messick's (1984, 1994, 1995) framework of construct validity, an older and well-know classical framework extensively used to evaluate tests. Goals of this comparison are to investigate whether quality criteria for competence assessments should be fundamentally different from traditional criteria such as Messick's and whether complementary quality criteria are needed that do more justice to the specific characteristics of assessment in competence-based education.

This article starts with our definition of competences, assessment of competences and introduces the idea of Competence Assessment Programmes or CAPs. Then, the ten quality criteria for competence assessment are described, followed by a short description of Messick's framework of construct validity. Finally, the two frameworks are compared and analogies and differences are formulated.

What Are Competences?

Before turning to matters related to the assessment of competences, the concept "competence" needs to be defined as accurately as possible, or at least an agreement must be

reached on a general description of the concept (i.e., determine a stipulative definition). The importance of defining the concept of competence appears from the fact that curricula and assessments are to a great extent determined by the learning outcomes we want students to achieve, which are in turn influenced by our conceptions of competence (Lizzio & Wilson, 2004).

The concept of competence is defined in many different ways (e.g., Eraut, 1994; Eraut, Alderton, Cole, and Senker, 1998; Lizzio & Wilson, 2004; Messick, 1984; Miller, 1990; Parry, 1996; Spencer & Spencer, 1993; Taconis, Van der Plas, & Van der Sanden, 2004; Tillema et al., 2000). A common notion of most descriptions of competence is that it consists of connected pieces of knowledge, skills and attitudes that can be used to adequately solve a problem. For example, Lizzio and Wilson (2004) see competence as the capacity to enact specific combinations of knowledge, skills and attitudes in appropriate job contexts. Taconis et al. (2004) stress that competence-based curricula should address knowledge, skills and attitudes in an integrated way, since each of these separately is not sufficient for the desired competent professional behaviour. Eraut (1994) also stresses not to regard skills as something separate from knowledge, as this would restrict the meaning of knowledge to propositional knowledge (i.e. propositions about skills, for example how to ride a bicycle) and exclude the practical know-how to perform these operations.

Eraut (1994) gives an elaborate overview of the development of the concept of competence, describing how research into the concept of competence has followed three main traditions. First, within the tradition of behaviourist psychology, very detailed specifications of competent behaviour have been produced, focusing purely on the technical process of task analysis but thereby neglecting the social and political dimensions of the development of competence. Second, generic approaches to competence aimed to identify overarching qualities linked to excellent job performance, and focused more on selection than on training or educational purposes. Spencer and Spencer's (1993) definition of competence, which focuses on the underlying characteristics of an individual, and causally relates these characteristics to

behaviour and performance, can be placed within this tradition. The third approach is based on cognitive constructs of competence and stems from cognitive psychological traditions. Researchers in this tradition sought to distinguish between competence and performance. For example, Chomsky (2006) made a distinction between linguistic competence and linguistic performance, which has implications for assessment methodology. Messick (1984), too, views competence as what a person knows and can do under ideal circumstances, whereas performance refers to what is actually done under existing circumstances.

Altogether, two aspects seem to be generally included in the definition of competence: the integration of knowledge, skills and attitudes, and a reference to a certain job context or job situation (e.g., Lizzio & Wilson, 2004; Parry, 1996). Eraut (1994) describes a similar dimension called scope, which concerns what a person is competent in, that is, the range of roles, tasks or situations for which a competence has been established or can be generalised to. This article uses the definition given by Eraut et al. (1998), as it incorporated these two aspects. They define competence by describing competent professional behaviour within a range of relevant job situations and the knowledge, skills and attitudes it requires.

Assessment of Competence

The past ten years in research on assessment have seen major changes, which even caused some authors to speak of a paradigm shift or a transition from a testing culture towards an assessment culture (e.g., Birenbaum, 1996; Birenbaum et al., 2006; Dierick & Dochy, 2001; Stiggins, 1991). A number of characteristics are usually used to describe the testing culture and the assessment culture, in which the two are presented as extremes along a continuum. The next section describes the characteristics usually attributed to the testing and assessment cultures by different authors. After this, the two cultures are taken together again, and our notion of assessment is presented.

The testing culture is often described as being based on a behaviouristic approach to instruction and learning, in which the learner is generally viewed as a passive receiver of the knowledge presented by the teacher. It goes back to the 1920s and the emerging industrial

society, when mass produced, efficient and cheap tests were needed to detect individual differences in achievement (Stiggins, 1991). Testing is seen as mainly addressing basic lower level skills and cognitive competences, based on the repetition of what has been taught in class or read in the textbook and is used almost exclusively summatively (Birenbaum, 1996, 2003). Assessment and instruction are separated from each other in such a way that teachers do the teaching and external measurement experts develop the assessment tools to be used by the teacher (Stiggins, 1991). The most common measurement format belonging to the testing culture is the choice response format, for example multiple choice, true/false, or matching items, administered through paper-and-pencil tests taken in class under time constraints and without allowance of the use of helping materials or tools. Only the product is evaluated; the process towards this end product is not taken into account in the final test result. The development of the tests and the criteria on which the students are judged remain unknown to the students. Regarding assessment quality, the testing culture relies on psychometric models of test development, scoring and interpretation of test results (Birenbaum, 1996). The psychometric approach stems from psychological research and the measurement of fixed traits (e.g. intelligence), based on which learners and their (potential) performances were distinguished. It is guided by the demand for objectivity and fairness in testing, requiring high levels of standardisation because of the high-stakes nature of tests within the testing culture.

The assessment culture arose from the growing criticism on traditional testing methods relating to the unrealistic nature of the tests, the loss of faith in them as valid measures of learning, and an over-reliance on tests as the ultimate goal of the instruction process (McDowell, 1995). Stiggins (1991) describes how the assessment culture started to emerge when US schools were held accountable for their educational outcomes and as a consequence started to realise that the majority of educational outcomes cannot be assessed by paper-and-pencil tests. The assessment culture is based on cognitive learning theories, in which learning is thought of as active construction of schemes in order to understand the material (Birenbaum, 1996). The student is an active participant, who shares responsibility for the learning process, practices self-

evaluation and reflection, and collaborates with the teacher and other students. Multiple forms of assessment are used, which are generally less standardised than the formats used in the testing culture. Often, the assessments are carried out without time pressure and using the tools or other helping materials that are also used in real life. The assessment tasks are meant to be interesting and authentic to students, and to engage them in meaningful learning processes. Both the product and the process are being assessed, and students reflect on and document their development in, for example, a portfolio. Assessment is not only used in a summative way, but also to guide the learner by providing feedback on the product and the process. Criteria are shared or even developed together with the students (Birenbaum, 1996; Dierick & Dochy, 2001). Regarding the quality of assessment, the psychometric approach is criticised for not fully capturing the unique nature of new assessments (Moss, 1994). The assessment culture rejects the fundamental belief that there can be universality of meaning as to what any grade or score represents and that it is possible to separate the goals of education from the means for their attainment (Berlak et al., 1992). The objects of measurement are not unchangeable personal traits, but the competence development of the student, which is even assumed to change over time. Therefore, it is argued that a new system of evaluating the quality of new assessments is needed, establishing a new school of edumetrics instead of psychometrics (Dierick & Dochy, 2001).

This article uses the term assessment to refer to all methods that can be used to determine and judge a learner's competences, including both traditional tests originating in the testing culture and new assessment methods stemming from the assessment culture, and including both formative and summative assessments. Our definition of assessment thus explicitly includes traditional tests. Cizek (1997) presents a definition of assessment that captures our view of assessment as a continuous process of assessing a learner's progress throughout (and beyond) education:

> (1) the planned process of gathering and synthesizing information relevant to the purposes of (a) discovering and documenting students' strengths and weaknesses, (b) planning and enhancing instruction, or (c) evaluating progress and making decisions

> about students, (2) the process, instrument or method used to gather the information. (p. 10)

This definition stresses the possibility to use assessment to guide and evaluate learner development and to enhance the quality of instruction, and also includes the use of assessment in a summative way, using it to certify learners. Second, it does not specify that assessment only comprises new forms of assessment, and can thus include traditional testing as well. This definition captures our views of assessment, to which we add the new idea of using programmes of assessment instead of single methods, which is elaborated on in the next section.

Competences Assessment Programmes

During the transition towards an assessment culture, which is still underway, a large number of new and different assessment methods have emerged. In assessment literature, many different names are used for these assessment methods such as performance assessment, competence assessment, direct assessment, authentic assessment, innovative assessment, continuous assessment, et cetera (e.g., Gulikers, Bastiaens, & Kirschner, 2004; Hambleton, 1996; Kniveton, 1996; McDowell, 1995). These methods together are often referred to as alternative assessment, because of their common background as being alternatives to traditional testing (e.g., Birenbaum, 1996; Dierick & Dochy, 2001; Maclellan, 2004). This name, however, might not have been very well chosen because it implies that newer assessment forms are a replacement of, or an alternative to older forms (Cizek, 1997). As such, we reject the term “alternative”.

This article considers both traditional tests and newer forms of assessment as necessary components of a Competences Assessment Programme or CAP. As such, CAPs combine elements of the testing culture and the assessment culture. Newer forms of assessment are not regarded as alternative to traditional tests, but as complementary to them. Traditional tests and newer assessments can be viewed as playing complementary rather than contradictory roles, although they are often presented as stemming from two contradictory cultures (Birenbaum, 1996). It is important to start thinking in terms of programmes of assessment because competences are such complex wholes of knowledge, skills and attitudes, that it is often argued

that one single assessment method is not enough to assess competences and that a mix of methods should be used instead (e.g. Chester, 2003; Dierick & Dochy, 2001; Maclellan, 2004; Stiggins, 1991; Van der Vleuten & Schuwirth, 2005). Maclellan (2004) warns that it would be unwise to assume that alternative assessments are the panacea for all assessment problems and Stiggins (1991) states that the challenge is to align the various assessment options we have to cover the broad array of achievement target we value. Dierick and Dochy (2001) write that traditional tests can be useful for certain purposes and that balanced and pluralistic assessment programmes should be used.

A CAP can thus be defined as a combination of both traditional and new forms of assessment, in which the actual combination of assessment methods used depends on the goals of the educational programme. No exact combination of forms of assessment can be given that unarguably or irrefutably defines a CAP, as the contents of a CAP depend on the competences being assessed and the breadth of the educational programme (i.e., a specific course, a semester, a school year, etcetera). A CAP as a whole should cover all educational goals, which, in competence-based education, implies that knowledge, skills and attitudes should be assessed in an integrated way. This already implies that the use of traditional tests alone is not sufficient in competence-based education. To give one example, a school involved in another CAP-related study (Baartman, Prins, Kirschner, & Van der Vleuten, in press), uses project-based education as a form of competence-based education. To assess students, they use a combination of written tests (both multiple choice and open questions), assessment of the products made during the projects, an assessment interview, and observations carried out by the teachers during classroom or practical work.

Quality Criteria for Competences Assessment Programmes

Now that competences and Competences Assessment Programmes have been defined, the question of how to guarantee the quality of these CAPs comes into play. The quality of CAPs in competence-based education cannot be ignored or undervalued, because high-stake decisions about learners are based on the outcomes of a CAP. The quality of traditional tests is generally

determined by quality criteria such as validity and reliability, but the question arises as to whether these criteria are sufficient to determine the quality of CAPs. Linn, Baker and Dunbar (1991) posit, for example, that it is critical to expand the criteria used to judge the adequacy of assessments now the forms of assessment we use are expanding. Benett (1993) argues to interpret the concepts of classical test theory (i.e., validity and reliability) in a broader sense, while retaining the essence of their meaning, and to search for appropriate ways of applying them to more qualitative assessment methods. Similarly, Martin (1997) states that as the notions of adequate assessments in competence-based education change, the notions of validity and reliability should change accordingly, without denying that new forms assessment should still be valid and reliable.

Besides broadening or altering the traditional notions of validity and reliability, different and other quality criteria have been proposed with the rise of the assessment culture (e.g., Dierick & Dochy, 2001; Linn et al., 1991). Again, here the question arises as to whether these new quality criteria should be considered as alternatives to the traditional criteria of validity and reliability or as complementary to them. As described in the previous section, this article considers new forms of assessment not as alternative, but as additional or enriching to traditional tests. In the same way that traditional tests should not be discarded for use in CAPs, traditional measures of reliability and validity are not fundamentally incorrect for determining the quality of a CAP, but are not sufficient to all of the aspects of a CAP. Therefore, this article argues that the traditional notions of validity and reliability need to be adapted for an adequate and fit for purpose judgment of the quality of CAPs in competence-based education, which is further elaborated on below. Besides this, new quality criteria derived from the assessment culture should be added to complement and enrich the traditional measures.

New Applications of Traditional Quality Criteria

In a previous study (Baartman et al., 2006) we already described a number of problems with regard to the use of validity and reliability to evaluate CAPs. Here, we further elaborate on those problems and describe how traditional notions of reliability and validity were adapted for

our framework of quality criteria for CAPs.

Reliability is concerned with the degree to which the same results would be obtained on a different occasion, in a different context, of by a different assessor. In classical test theory, reliability is about the accuracy of measurement, operationalised as for example test-retest comparisons or split-half methods. The goal of classical test theory is to discriminate between students (Martin, 1997). Assessment in competence-based education, however, is not about discriminating between students or comparing students to each other (norm-referenced assessment), but about the decision whether or not a student is competent or not (criterion-referenced assessment). Cronbach, Linn, Brennan and Haertel (1997) describe a number of other features in which new assessments differ from traditional testing, making psychometric approaches and criteria inappropriate, such as the fact that tasks are often complex and open ended and that decisions are based on unconventional combinations of scores and assessor judgments. Contrary to traditional testing, assessing competence always involves a domain expert's judgment and the main doubts regarding the reliability of competence assessment pertain just this reliance on human subjective judgments. Thus, reliability is often phrased in terms of the agreement of judges or interrater reliability. This aspect of reliability, however, is not the only important component. Not only consistency across raters is needed, but also across tasks that vary in content or format (Dunbar, Koretz, & Hoover, 1991). Studies applying generalizability theory have shown that reliability across judges is far greater than reliability across tasks, due to the interaction between the student and the assessment task. Each assessment task calls on different skills and motivations on which certain students are strong and others weak (Cronbach et al., 1997). These studies also showed that acceptable levels of reliability across judges can be reached in any assessment format, provided that multiple assessments are used, which on its turn shows that reliability is not conditional on objectivity and standardisation (Van der Vleuten & Schuwirth, 2005).

Concluding, the essence of the meaning of reliability, that is consistency of results across occasions, contexts and assessors, can be retained for CAPs, but due to the different nature of

CAPs, the actual applications have to be altered. First, generalizability studies have shown that multiple assessments should be used to reach acceptable levels of reliability across judges and across methods (Wass, McGibbon, & Van der Vleuten, 2001). This aspect of reliability is included in our framework under the criterion "reproducibility of decisions", which describes that the final (high-stake) decisions about students should be based on multiple assessors, multiple occasions, multiple contexts, and multiple methods. Second, Benett (1993) noted that assessments carried out in less controllable and standardised contexts, such as assessment in the workplace, can nonetheless be based on a set of tasks, which, although not identical, are consistent with respect to key features of interest (e.g. a common assessment procedure, a theme and purpose). This aspect of reliability is included in our framework under the criterion "comparability". Both quality criteria are further described in the next section.

With regard to validity, a major problem lies in the fact that many different definitions of validity are being used. The Standards for Educational and Psychological Testing (American Educational Research Association, American Psychological Association, & National Council on Measurement, 1999) defined validity as: "the degree to which evidence and theory support the interpretations of test scores entailed by proposed uses of tests" (p.9). Although few would dispute this definition of validity or ignore its importance, the actual criteria for examining validity vary widely (Miller & Linn, 2000). Just to mention a few examples, Kane (2004, p. 135) describes validity as: "Do the scores yielded by the procedure supply the kind of information that is of interest, and are these scores helpful in making good decisions? Validity addresses these two questions … ". Benett (1993, p. 83) defines validity as "what it is that is being assessed … the intention of the assessor and the nature of what is to be assessed". Kane (2001) presents an elaborate overview of the development of the validity concept throughout the 20th century, in which he describes the development of the concept from a criterion-based model ("how well does the test score predict the criterion score") to a construct model ("construct validity as a unified framework for validity"). Describing current conceptions of validity, he argues for an argument-based approach to validity, which entails an analysis of all evidence for and against the

proposed interpretation of the test scores and an evaluation of the plausibility of these arguments. A final well-know framework of validity not included in Kane's (2001) overview is Messick's (1994, 1995) framework of construct validity. Messick sees construct validity as a unified and overarching validity concept, but nonetheless distinguishes six aspects of construct validity: content validity, substantial validity, structural validity, consequential validity, external validity and generalisability. Messick introduced the idea of consequential validity, thereby broadening the validity concept to include the consequences of test use and test scores for students.

Concluding, many different definitions of validity are being used. Both the breadth and the complexity of the concept make it difficult to work with in practice (Crooks & Kane, 1996) and it is difficult to disentangle its many intertwined facets (Birenbaum, 1996). According to Crooks and Kane (1996), new approaches are needed that help us organise our thinking about important validation questions. Two approaches can be distinguished that aim at this clarification. First, researchers like Kane (2001) and Shepard (1993) argue for an argument-based approach to validity. Though it may be a very valuable approach to collect sources of evidence to demonstrate validity, no clear definition is given of the concept of validity itself, leaving it unclear to practitioners what the evidence should be collected for. Second, sets of quality criteria are being identified that have proven helpful in identifying the issues that deserve attention in validation, and that clarify how specific assessment concerns relate to the more global issues of construct validity (Crooks & Kane, 1996). Examples of this approach can be found in the work of Linn et al. (1991) and in Messick's aspects of construct validity (e.g., 1994, 1995). Our framework of quality criteria for CAPs can also be placed within this approach. The concept of validity needs to be clarified and further operationalised for practical use. Validity is not just a matter of assessing the right constructs, but increasingly pertains the actual and correct use of assessment instruments. This also implies that practitioners or users, who work with the assessment instruments in practice, to a great extent determine the quality of the assessments. They need to be able to understand and work with concepts like validity, and thus the development and validation of assessment instruments cannot be separated from its context.

To recapitulate, the goal of this article is to provide a framework of quality criteria for CAPs that is both consistent with current theoretical understandings of reliability and validity, and the nature and potential of new forms of assessment. The essence of the meaning of reliability and validity is incorporated in our framework, but they are applied in a different way. The essence of reliability is subsumed under the quality criteria comparability and reproducibility of decisions. The link between the quality criteria comprising the framework and the concept of validity are further elaborated on by means of a comparison between our framework and Messick's aspects of construct validity.

Ten quality criteria for CAPs

The framework of quality criteria presented in this article is based on a literature review and is a synthesis of work by many different authors (e.g., Alderson & Wall, 1993; Bachman, 2002; Brown, 2004; Benett, 1993; Birenbaum, 1996; Cronbach et al., 1997; Crooks & Kane, 1996; Dunbar et al., 1991; Frederiksen & Collins, 1989; Gulikers et al., 2004; Haertel, 1991; Hambleton, 1996; Kane, 1992, 2004; Linn et al., 1991; Martin, 1997; McDowell, 1995; Prodromou, 1995; Uhlenbeck, 2002; Van der Vleuten & Schuwirth, 2005). The interested reader is referred to previous publications on the ten quality criteria (Baartman et al., 2006, in press). Here, we focus on the terminology and definitions used by other authors to define quality criteria for assessment. These definitions were compared, and one term was chosen for this research. The goal of the framework is to provide a clear definition of all criteria that can be used to evaluate a CAP, so as to enable further operationalisation in an instrument in further studies. The ideas of reliability and validity are incorporated in the framework, but worked out in a different way in the framework. In addition, new quality criteria derived from the assessment culture are also included. Together, they provide an integral framework of quality criteria for CAPs.

Authenticity as a quality criterion for assessment is generally described as the degree of resemblance of a CAP to the criterion situation, meaning that an assessment should reflect the competences needed in the future workplace (Bachman, 2002; Brown, 2004; Dierick & Dochy, 2001; Uhlenbeck, 2002). Gulikers et al. (2004) elaborate on the concept of authenticity and

distinguish five dimensions that can vary in authenticity.

Cognitive complexity resembles authenticity in that it also relates to the future professional life, but it focuses on the fact that CAPs should also reflect higher cognitive skills (Dierick & Dochy, 2001; Hambleton, 1996; Linn et al., 1991). Bachman (2002) describes a related concept called interactiveness, which is defined as the extent to which the test tasks engage the processes and strategies that are part of the construct being assessed. The use of performance assessments, however is no guarantee that higher cognitive skills are indeed being measured (Hambleton, 1996). To gain insight into the thinking processes applied by students, MacLellan (2004) suggests having them provide a rationale for their answer or action chosen.

Fairness and related concepts are described by a number of authors. Brown (2004) mentions equal opportunities as an important quality criterion, noting that all participants need to be given the opportunity to demonstrate their abilities and maximise their potential. Dierick and Dochy (2001) note that bias can arise from assessment tasks that are not adjusted to the educational level of the learners or that contain cultural aspects not familiar to all learners. Related to this is the scope or coverage of the assessment (Frederiksen & Collins, 1989; Linn et al., 1991; Uhlenbeck, 2002) implying that the tests should cover all knowledge, skills and strategies required to do well. An assessment should, thus, reflect the knowledge, skills and attitudes of the competences at stake, excluding irrelevant variance (see also: Haertel, 1991; Hambleton, 1996).

Meaningfulness entails that a CAP should have a significant value for both teachers and learners (Linn et al., 1991), to which the importance in the eyes of employers could be added. The assessment should get students to deal with meaningful problems that provide worthwhile educational experiences. McDowell (1995) stressed that for learners to perceive an assessment as meaningful, they need to perceive a link between the assessment task and their personal interests. Meaningfulness, thus, is different from authenticity as an assessment that is authentic for an experienced practitioner might not be meaningful to a novice (Gulikers et al., 2004).

Directness is the degree to which teachers or assessors can immediately interpret the

assessment results, without translation from theory into practice (Dierick & Dochy, 2001; Frederiksen and Collins, 1989). Frederiksen and Collins note that "any indirectness in the measure will lead to a misdirection of learning effort by test takers" (p. 30). Linn et al. (2001) write that direct assessments of performance appear to have the potential of enhancing validity.

Transparency means that a CAP must be clear and understandable to all participants (Brown, 2004; Dierick & Dochy, 2001; Frederiksen & Collins, 1989). Learners should know the scoring criteria, who the assessors are and what the purpose of the assessment is. They should know what is expected of them so as to be able to prepare for the assessment and adjust their learning processes accordingly (Frederiksen & Collins, 1989). This is also true for teachers and/or assessors, who should know and understand the entire CAP to be prepared for their role as assessor (Baume, Yorke, & Coffey, 2004). As an indication of transparency, Hambleton (1996) suggests to check whether learners can judge themselves as accurately as trained assessors.

Educational consequences pertains the effects the CAP has on learning and instruction (Dierick & Dochy, 2001; Linn et al., 1991; Uhlenbeck, 2002, Van der Vleuten & Schuwirth, 2005). A collection of evidence is needed about the intended and unintended, positive and negative effects of the assessment on how teachers and learners view the goals of education and adjust their learning and teaching activities accordingly. For summative purposes, unintended factors and adverse impact are especially important. This criterion is also related to effects like backwash (Prodromou, 1995) or washback (Alderson & Wall, 1993).

Reproducibility of decisions is the term chosen here to address the fact that (high-stakes) decisions made about students should be based on multiple assessments, carried out by multiple assessors and on multiple occasions. Different terms are used by different authors. Bachman (2002) uses generalisability and extrapolation and Linn et al. (1991) use transfer and generalisability to refer to the degree to which assessment results can be generalised to broader student domains. The purpose of an assessment is not a performance in one specific situation, observed by one assessor, but should enable the assessor to draw more general conclusions about a learner's competences. Reproducibility includes the idea of human judgment and the necessity

of adequate sampling of tasks. In other words, reproducibility is determined for the final decisions made, and not for single assessment tasks. A CAP, which includes multiple assessment forms, implies looking at the overall reproducibility of the CAP as a whole.

Comparability addresses the fact that a CAP should be conducted in a consistent and responsible way. Uhlenbeck (2002) relates comparability to the fact that the conditions under which the assessment is carried out should, as much as possible, be the same for all learners and scoring should occur in a consistent way. Because assessment methods in CAPs are generally less standardised than classical tests, this necessitates reliance on human judgment. Therefore, the consistency of the scoring procedure is very important (Haertel, 1991). Benett (1993) notes that comparability can be achieved when the assessment is based on a set of tasks which, though not identical, are consistent with respect to key features of interest.

Finally, costs and efficiency is addressed by Brown (2004) as efficiency, by Linn et al. (1991) as cost and efficiency and by Uhlenbeck (2002) as practicability. This criterion is especially important when undertaking competence assessment, because of the complexity of undertaking such an assessment. Assessment choices are not only influenced by educational factors, but also by financial, managerial and institutional ones. Learners should find the assessment tasks manageable (Brown, 2004) and evidence needs to be found that the additional investments in time and resources are justified by the positive effects of competence assessment, such as improvements in learning and teaching (Hambleton, 1996).

The ten quality criteria for CAPs are depicted on the left side of Figure 1. In the following section, Messick's (1994, 1995) framework of construct validity is shortly described, after which our framework of quality criteria is compared to it. The goal of this comparison is twofold. First, it was determined whether there really is a fundamental difference between classical and new quality criteria. As our framework comprises both classical and new quality criteria, we expected a partial overlap between the two frameworks. Second, it was investigated whether the framework of quality criteria for CAPs does better justice to assessments in competence-based education. We expected some quality issues which are specifically important

in competence-based education to be missing in Messick's framework. Messick (1994) stated that "Validity criteria especially tailored for performance assessment … are for the most part consistent with but less extensive than the general validity standards [i.e. Messick's] (p.13). The goal of this article is to show that this relationship is reversed: quality criteria for competence-based education match with Messick's general aspects of validity, but are more extensive and add some quality aspects not accounted for in Messick's framework.

- Insert Figure 1 about here -

A Psychometric Validity Framework

Messick's (1984, 1994, 1995) framework of construct validity is depicted on the right-hand side of Figure 1. This framework was chosen as an example of the psychometric tradition, since it covers the whole breadth of psychometric quality control issues for testing and assessment. Messick integrated the three traditional aspects of validity (content, construct and criterion) into one criterion called construct validity, and included the idea of consequential validity, the effect assessment has on education. In his framework, Messick describes six aspects of construct validity, namely: content, substance, structure, consequences, externality, and generalisability. The content aspect prescribes that an assessment task for competence assessment should encompass the knowledge, skills and attitudes comprising the competence. The substantive aspect, sometimes called the syntactic aspect (Frederiksen & Collins, 1989), adds the need for the thinking processes used during the assessment to be a reflection of the processes used by practitioners in the construct field. The structural aspect of construct validity concerns the fidelity of the scoring structure used for the assessment programme, which should be consistent with what is known about the structure of the construct domain (i.e. the competence). The consequential aspect relates to the positive and negative, intended and unintended consequences of the assessment procedure, with regard to use of the assessment for certifying and the effect the assessment possibly has on learning and teaching. The generalisability aspect describes the correlation with other tests representing the construct or parts of it, determined across time, occasions and observers. The question that has to be asked in

this respect is whether the interpretation of the score that was based on one task can be generalised to other domain-specific tasks. Finally, the external aspect of construct validity applies to the relationship between scores obtained in the assessment and other measures of the same construct and other constructs. In this respect, scores on the assessment that represent the construct should show high correlations with other construct-relevant measures, while low correlations should be found with construct-irrelevant measures.

Comparing the Ten Quality Criteria to Messick's Framework

This section compares the ten quality criteria for CAPs to Messick's (1984, 1994, 1995) framework of construct validity. Per quality criterion, the analogies and differences are described. The differences pertain quality aspects that are missing in Messick's framework, but which are important for assessments in competence-based education. Here, our aim is not to deny the importance of Messick's framework or to replace it, but rather to look for quality elements that are important for specifically assessing competences, an aim not included in psychometric frameworks. Table 1 summarises the analogies and the differences between the ten quality criteria and Messick's framework.

- Insert Table 1 about here -

Authenticity implies that the tasks used in a CAP should reflect the type of tasks that can be encountered in an occupational area and should be as realistic as possible. Authenticity is mainly related to the content aspect of validity, described by Messick (1994, 1995) as the fact that an assessment should include all knowledge, skills and attitudes the competence comprises. A way of including all necessary knowledge, skills and attitudes in an assessment, is to reflect the job situation as accurately as possible, thus making the assessment more authentic. Both criteria do not match perfectly, though. The inclusion of all necessary knowledge, skills and attitudes is not enough for a CAP to be authentic. Knowledge, skills and attitudes have to be assessed in an integrated way, as they are used as an integrated whole in a job situation. Second, authenticity comprises more than content validity. It also includes the work environment and the social context, which have to reflect the future job situation as well. These aspects of authenticity

are important for assessing competences, but are not included in Messick's framework.

Cognitive complexity mainly resembles the substantive aspect described by Messick (1994, 1995). Substantiveness is defined as the degree to which the thinking processes employed during the assessment reflect those used by practitioners in the field. Both cognitive complexity and the substantive aspect of construct validity described by Messick require an analysis of the thinking processes used when solving the assessment problem. The tasks used in the CAP should reflect these processes. The difference between cognitive complexity and Messick's substantive aspect lies in the fact that Messick focuses on technical analyses of the assessment tasks by means of for example task analysis and think aloud protocols. Although these procedures are very valuable, a different operationalisation of cognitive complexity can be added, namely the assurance of cognitive complexity during the assessment itself, for example by asking learners to explain their choices during a performance assessment.

Fairness is linked to the content aspect. To give learners a fair chance of demonstrating their competences, the assessment tasks in a CAP should be varied to cover the entire domain of the construct or competences. Fairness as a criterion for CAPs is also linked to the structural aspect because the scoring criteria used in the assessment procedure should not show any bias to certain groups of learners. One step towards achieving fairness is taken when the scoring criteria follow the structure of the construct (i.e. the competence) and the weights used for the scoring criteria are adjusted accordingly. Though analogies can be found, Messick's aspects again seem to focus mainly on the technical relationships between assessment content and structure and the construct's. Fairness in our view should not only focus on covering the domain, but also on recognising individual differences between learners and assuring suitability of the assessment for the entire learner population, with regard to style and rate of learning. Fairness is typically something human. It includes the stakeholders in the assessment process. With regard to assessors: are they biased towards certain (groups of) learners. With regard to learners (and all other stakeholders): do they perceive the assessment as fair and to they have possibilities to appeal a decision made.

Meaningfulness is related to the content aspect because the assessment tasks should be recognisable to learners and considered valuable by them. It is also related to the substantive aspect because the assessment must contain recognisable behaviour and processes needed in the workplace and the assessment should be considered worth to do by learners. Finally, meaningfulness is related to the consequential aspect of validity described by Messick (1995) as the positive and negative consequences of the assessment procedure. For the assessment in itself to be a valuable learning experience and a guidance of learning processes, it should be recognisable and valued by the learner. If this is not the case, the expected or desired consequences of an assessment may fail to occur. Our comparison shows that analogies between meaningfulness and content and consequential validity can be found, but these analogies mainly arise from results of other studies. For example, McDowell (1995) stated that the assessment tasks (content) have to be meaningful to learners in order to achieve an effect on learning (consequences). Again, in Messick's framework quality is mainly determined by looking at the components of an assessment and their relationship to the construct being measured. Since the quality of assessment in competence-based education is largely determined by the people who carry out the assessment, an important element that needs to be added to Messick's framework is the stakeholders in the assessment process (learners, assessors, employers) who have to experience the assessment as meaningful. For example, the learners need meaningful feedback and assessment criteria to guide their learning process.

Directness is an important criterion for CAPs because of the presumed effect it has on learning and teaching, arguing that indirect performance measures may distort the focus of teaching and learning (Linn et al., 1991). Comparing this criterion to Messick's (1994, 1995) framework, directness mainly falls under the head of consequential aspect of validity. The way in which learners are assessed, in a direct way through for example performance assessment, or in an indirect way through interpretation of (written) answers, affects the instructional process and the way of learning. Directness differs from consequential validity in that it not only focuses on the assessment method itself that has an effect on learning. It also includes the fact that the

results of a direct assessment are easier to interpret for assessors who have to judge if a student is competent in handling complex, uncertain job situations.

Transparency is related to the structural aspect of construct validity, which means that the structure of the scoring system should be consistent with what is important and less important in the structure of the competence. Transparency is linked to the structural aspect because the scoring criteria and the weights used should be clear to learners (Messick, 1994). Transparency is also related to the consequential aspect of construct validity described as the influence of assessment on the learning process (Messick, 1995). The criteria should be known and clear to learners, because this improves learning (Dochy & McDowell, 1997). On the other hand, like it was the case for meaningfulness, the links between transparency and structural and consequential validity are made using other theories. For example, transparency is related to the scoring structure, because Gibbs (1999) showed that it is important to assure the scoring criteria are clear to learners to achieve an effect on learning. Messick does not specifically state that the scoring system must be transparent in order to assure other effects. He focuses on the more technical test aspects and only states that it must cover and represent the structure of the construct being measured.

Educational consequences refers to the effects the assessment has on the learning process and the design of the educational environment. Since one of the effects of an assessment should be stimulating competence acquisition (as a positive consequence), educational consequences as a quality criterion is clearly linked to the consequential aspect of validity described by Messick (1994, 1995). Although the two concepts are fairly similar, Messick's consequential aspect describes the effect an assessment has on learning in both positive and negative terms. Educational consequences, as a quality criterion in our framework, specifies a CAP must have a positive effect on student learning, as this is one of the major goals of formative and also summative assessment. We view assessment as part of the learning process, and not just as a measurement at the end of it. Secondly, Messick evaluates the positive and negative consequences after the assessment has taken place. In our view, if we really want to use

assessment to stimulate learning, the impact on learning should be purposefully used as a guiding principle when choosing different assessment forms.

Reproducibility of decisions has been described by a number of authors as being related to generalisability, defined by Messick (1994) as the question whether the outcomes of an assessment can be applied to other populations, settings and tasks. Generalisability is increased when a larger sample across content and situations is used. The difference between reproducibility and generalisability lies in the fact that Messick compares different tests assuming to measure the same thing to achieve generalisability. This implies something like a true test or a true score exists to which newer tests can be compared. In competence-based education, reproducibility of decisions can be achieved by combining different information sources in a CAP (e.g. assessors, tasks, situations) to get a better and more complete picture of a learner's competences. The idea is that assessing in for example a number of different situations make it more likely that the same decision about a learner is made in again another situation (reproducibility), which makes the results more generalisable.

Comparability is related to the external aspect, which pertains the correlations of the assessment with multitrait-multimethod comparisons (Messick, 1994, 1995). Actually, the external aspect is prerequisite for generalisability; when the relationships between the assessment and other measures reflect the competence, generalisation is possible. In the same way, comparability makes reproducibility of decisions easier to achieve. When two assessments of the same competence, taken at different times or by different observers, are highly comparable and show high correlations (comparability or external aspect), it is likely that the decisions based on the outcomes of these two measures will be reproducible by different observers and in different situations (reproducibility of decisions or generalisability). On the other hand, comparability is not exactly the same as Messick's external aspect. Whereas Messick mainly compares different tests measuring the same and different constructs (looking for high and low correlations), comparability focuses on the conditions under which an assessment is carried out, for example if tasks, scoring criteria and circumstances are comparable across different assessments.

Costs and efficiency is difficult to link to Messick's framework of validity, because his framework does not explicitly focus on implementation issues. Messick himself (1994) links cost and efficiency to the external aspect of validity: "Yet validity of performance tests should not be conceived of in terms of improved costs and efficiency alone, but rather in terms of costs and efficiency relative to the benefits, which is the general external validity criterion of utility" (p.21). He does not further elaborate on how utility is part of his external aspect of validity and the relationship seems somewhat arbitrary, though. This criterion is very important for CAPs, because it can never be successfully implemented if the costs are too high or it takes teachers, assessors and learners too much time. In this way, our framework broadens the idea of quality to include implementation issues. In our view, an assessment looses part of its quality if is not carried out well because of money and time constraints.

Conclusions and discussion

The goal of this article was to explore what quality criteria are needed to evaluate assessments in competence-based education. The idea of assessment programmes (CAPs) was introduced and a framework of ten quality criteria for CAPs was presented, including both criteria from traditional, psychometric traditions and criteria from the newer assessment culture, related to current ideas about competence-based education. A comparison was made between this framework and Messick's psychometric framework of construct validity to investigate (1) whether classical and new quality criteria are really fundamentally different, and (2) whether the 10-criterion framework does better justice to the specific characteristics of assessments in competence-based education.

With regard to the first question, a comparison of the two frameworks shows that many relationships exist between classical and new quality criteria (see also Figure 1). It also shows that, though related, they are operationalised in different ways and do not completely overlap (see Table 1). As predicted, our framework partly overlaps with Messick's, but also adds and elucidates quality aspects. First, as became apparent from Figure 1, the more traditional quality criteria in our framework (comparability, reproducibility, and transparency) mainly overlap with

Messick's structural and external aspects and with generalisability. The newer criteria in our framework tend to fall within two categories: authenticity and cognitive complexity are related to the content and substantive aspect, and meaningfulness, educational consequences and directness are mainly related to the consequential aspect of validity. This aspect is exactly the one that distinguishes Messick's framework from the traditional division into content, construct and criterion validity. It seems a logical finding that many of our newer criteria are mainly related to this "newest" aspect of Messick's framework. Second, the comparison shows that, for each criterion except for costs and efficiency, analogies between the two frameworks can be found. These analogies mainly comprise the fundamental, general ideas of quality of assessment. When it comes to operationalising the quality criteria, differences between the frameworks become apparent. The main differences seem to be that (1) Messick mainly focuses on the technical issues of test quality whereas our framework includes the stakeholders in the assessment process as important determinants of assessment quality, (2) Messick does not focus on the feasibility of carrying out assessments. To sum up, this comparison warrants the tentative conclusion that, on a fundamental level, analogies between classical and new quality criteria can be found, but when operationalised, the two traditions are clearly different.

Our second research question focused on whether the10-criterion framework does better justice to the specific characteristics of assessment in competence-based education. The comparison shows that our framework adds the aspect of costs and efficiency, which could not be related to Messick's framework. Costs and efficiency are especially important for implementing a CAP, while practical implementation issues are not clearly included in Messick's framework. Secondly, the newer competence-based quality criteria in our framework could be linked to Messick's, but they are more clearly distinguished and operationalised in our framework. Criteria like transparency and meaningfulness more explicitly include the learner and his or her point of view in quality control issues. Although Messick also focuses on the consequences for the learner, he does so from the point of view of a test developer ("we know what is good for the learner"). The starting point of our framework is to involve all stakeholders

(teacher, learner, industry) in the assessment procedure and to pay attention to all interests and opinions. This is one of the fundamental ideas of competence-based education. From the comparison conducted in this study, we can thus tentatively conclude that the 10-criterion framework for CAPs does better justice to assessments in competence-based education.

Some critical remarks about this study can be made. The tentative conclusions that have been made are based solely on a literature review and a theoretical comparison of two frameworks. Further, and empirical, research is needed to show whether the proposed framework needs to be adapted or complemented. Also, the opinions of stakeholders in the assessment process need to be investigated. Therefore, experts' and teachers' opinions on quality criteria for CAPs were investigated (Baartman et al., 2006; in press). Secondly, further research is needed into the practical use of the quality criteria to evaluate CAPs. The ideas of combining psychometric and edumetric quality criteria and the evaluation of programmes of assessment instead of single methods are new. The criteria included in the framework need to be further operationalised for practical use in educational institutes and evidence must be gathered to determine whether the criteria can actually be used to distinguish "good" and "bad" CAPs. Also, it is necessary to investigate when a CAP *as a whole* complies with quality criteria like the ones proposed here. For example, do all forms of assessment included in a CAP have to comply with all criteria (non-compensatory), or is it sufficient if one or two assessment forms comply with a criterion (compensatory)?

Concluding, the 10-criterion framework seems to add important quality aspects to Messick's (1994, 1995) framework, while not playing down the importance of his work and the influence of this psychometric framework. Competence assessment should build on psychometric work. The ten criteria for CAPs are a first step in this direction. Together, they highlight relevant quality issues for CAPs in competence-based education.

Acknowledgement

The authors would like to thank Dr. Frans Prins for his useful comments and suggestions. We also would like to thank the reviewers for their feedback on the original manuscript. Without their work, this would have been a lesser article.

References

Alderson, J. C., & Wall, D. (1993). Does washback exist? *Applied Linguistics, 14,* 115-129.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education (1999). *Standards for educational and psychological testing.* Washington, DC: AERA.

Baartman, L. K. J., Bastiaens, T. J., Kirschner, P. A., & Van der Vleuten, C. P. M. (2006). The wheel of competency assessment. Presenting quality criteria for Competency Assessment Programs. *Studies in Educational Evaluation, 32*, 153-177.

Baartman, L. K. J., Prins, F. J., Kirschner, P. A., & Van der Vleuten, C. P. M. (in press). Determining the quality of Competences Assessment Programmes: A Self-Evaluation Procedure. *Studies in Educational Evaluation*.

Bachman, L. F. (2002). Alternative interpretations of alternative assessments: Some validity issues in educational performance assessments. *Educational Measurement: Issues and Practice, 21*, 5-18.

Baume, D., Yorke, M., & Coffey, M. (2004). What is happening when we assess, and how can we use our understanding of this to improve assessment? *Assessment & Evaluation in Higher Education, 29*, 451-477.

Benett, Y. (1993). The validity and reliability of assessments and self-assessments of work-based learning. *Assessment & Evaluation in Higher Education, 18*, 83-95.

Berlak, H., Newman, F., Adams, E., Archbald, D., Burgess, T., Raven, J. & Romberg, T. (1992). *Towards a new science of educational testing and assessment*. New York, University of New York Press.

Biggs, J. (1996). Enhancing teaching through constructive alignment. *Higher Education, 32*, 347-364.

Birenbaum, M. (1996). Assessment 2000: Towards a pluralistic approach to assessment. In M. Birenbaum & F. J. R. C. Dochy (Eds.), *Alternatives in assessment of achievement, learning processes and prior knowledge.* (pp. 3-29). Boston: Kluwer Academic Publishers.

Birenbaum, M. (2003). New insights into learning and teaching and their implications for assessment. In M. Segers, F. J. R. C. Dochy, & E. Cascallar (Eds.), *Optimising new*

modes of assessment: In search of qualities and standards (pp. 13-36). Dordrecht, The Netherlands: Kluwer Adademic Publishers.

Birenbaum, M., Breuer, K., Cascallar, E., Dochy, F., Dori, Y., Ridgway, J., Wiesemes, R. (2006). EARLI position paper. A learning integrated assessment system. *Educational Research Review, 1,* 61-67.

Brown, S. (2004). Assessment for learning. *Learning and Teaching in Higher Education, 1*, 81-89.

Chester, M. D. (2003). Multiple measures and high-stakes decisions: A framework for combining measures. *Educational Measurement: Issues and Practice, 22*, 32-41.

Chomsky, N. (2006). *Language and mind* (3rd ed.). Cambridge, Great Britain: Cambridge University Press.

Cizek, G. J. (1997). Learning, achievement, and assessment: Constructs at a crossroads. In G. D. Phye (Ed.), *Handbook of classroom assessment: Learning, achievement, and adjustment.* (pp. 1-32). San Diego: Academic Press.

Cronbach, L. J., Linn, R. L., Brennan, R. L., & Haertel, E. H. (1997). Generalizability analysis for performance assessments of student achievement or school effectiveness. *Educational and Psychological Measurement, 57,* 373-399.

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

Dierick, S. & Dochy, F. J. R. C. (2001). New lines in edumetrics: new forms of assessment lead to new assessment criteria. *Studies in Educational Evaluation, 27*, 307-329.

Dochy, F. J. R. C., & McDowell, L. (1997). Introduction: Assessment as a tool for learning. *Studies in Educational Evaluation, 23,* 279-298.

Dunbar, S. B., Koretz, D. M., & Hoover, H. D. (1991). Quality control in the development and use of performance assessment. *Applied Measurement in Education, 4*, 289-303.

Eraut, M. (1994). *Developing professional knowledge and competence*. London: Routledge Falmer.

Eraut, M., Alderton, J., Cole, G., & Senker, P. (1998). *Development of knowledge and skills in employment*, Final report of a research project funded by The Learning Society Programme of the Economic and Social Research Council, Brighton, University of Sussex.

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

Frederiksen, N. (1984). The real test bias. Influences of testing on teaching and learning. *American Psychologist, 39,* 193-202.

Gibbs, G. (1999). Using assessment strategically to change the way in which students learn. In S. Brown & A. Glaser (Eds.), *Assessment matters in higher education* (pp. 41-53). Buckingham, United Kingdom: SRHE.

Gulikers, J. T. M., Bastiaens, T. J., & Kirschner, P. A. (2004). A five-dimensional framework for authentic assessment. *Educational Technology Research and Design, 53*, 67-87.

Haertel, E. H. (1991). New forms of teacher assessment. *Review of Research in Education, 17,* 3-29.

Hambleton, R. K. (1996). Advances in assessment models, methods, and practices. In D. C. Berliner & R. C. Calfee (Eds.), *Handbook of educational psychology* (pp. 899-925). New York: MacMillan.

Hambleton, R. K. & Murphy, E. (1992). A psychometric perspective on authentic measurement. *Applied Measurement in Education, 5*, 1-16.

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

Kane, M. T. (2004). Certification testing as an illustration of argument-based validation. *Measurement: Interdisciplinary Research and Perspectives, 2,* 135-170.

Kniveton, B. H. (1996). Student perceptions of assessment methods. *Assessment & Evaluation in Higher Education, 21,* 229-237.

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

Lizzio, A. & Wilson, K. (2004). Action learning in higher education: an investigation of its potential to develop professional capability. *Studies in Higher Education, 29,* 469-488.

Martin, S. (1997). Two models of educational assessment: A response from initial teacher education: If the cap fits … *Assessment & Evaluation in Higher Education, 22*, 337-342.

Maclellan, E. (2004). How convincing is alternative assessment for use in higher education? *Assessment & Evaluation in Higher Education, 29*, 311-321.

McDowell, L. (1995). The impact of innovative assessment on student learning. *Education and*

Training International, 32, 302-313.

Messick, S. (1984). The psychology of educational measurement. *Educational Measurement, 21*, 215-237.

Messick, S. (1994). The interplay of evidence and consequences in the validation of performance assessments. *Educational Researcher, 23,* 13-23.

Messick, S. (1995). Validity of psychological assessment. Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist, 50,* 741-749.

Miller, G. E. (1990). The assessment of clinical skills / competence / performance. *Academic Medicine, 65*, 63-67.

Miller, M. D., & Linn, R. L. (2000). Validation of performance-based assessments. *Applied Psychological Measurement, 24*, 367-378.

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

Parry, S. B. (1996). The quest for competences: Competence studies can help you make HR decisions, but the results are only as good as the study. *Training, 33*, 48-56.

Prodromou, L. (1995). The backwash effect: from testing to teaching. *ELT Journal, 49,* 13-25.

Schuwirth, L. W. T., & Van der Vleuten, C. P. M. (2004). Changing education, changing assessment, changing research? *Medical Education, 38*, 805-812.

Shepard, L. A. (1993). Evaluating test validity. *Review of Research in Education, 19*, 405-450.

Spencer, L. M., & Spencer, S. M. (1993). *Competence at work. Models for superior performance.* New York: John Wiley & Sons.

Stiggins, R. J. (1991). Facing the challenges of a new era of educational assessment. *Applied Measurement in Education, 4*, 263-273.

Stokking, K., Van der Schaaf, M., Jaspers, J. & Erkens, G. (2004). Teachers' assessment of students' research skills. *British Educational Research Journal, 30*, 93-116.

Taconis, R., Van der Plas, P., & Van der Sanden, J. (2004). The development of professional competencies by educational assistants in school-based teacher education. *European Journal of Teacher Education, 27*, 215-240.

Tillema, H. H., Kessels, J. W. M., & Meijers, F. (2000). Competences as building blocks for integrating assessment with instruction in vocational education: A case from The Netherlands. *Assessment & Evaluation in Higher Education, 25,* 265-278.

Uhlenbeck, A. M. (2002). *The development of an assessment procedures for beginning teachers of English as a foreign language.* Doctoral dissertation, University of Leiden, ICLON Graduate School of Education, The Netherlands.

Van der Vleuten, C. P. M., & Schuwirth, L. W. T. (2005). Assessing professional competence: from methods to programmes. *Medical Education, 39*, 309-317.

Wass, V., McGibbon, D., & Van der Vleuten, C. P. M. (2001). Composite undergraduate clinical examinations: How should the components be combined to maximize reliability? *Medical Eduation, 35*, 326-330.

Table 1. Analogies and differences between the 10-criterion framework and Messick's aspects of construct validity

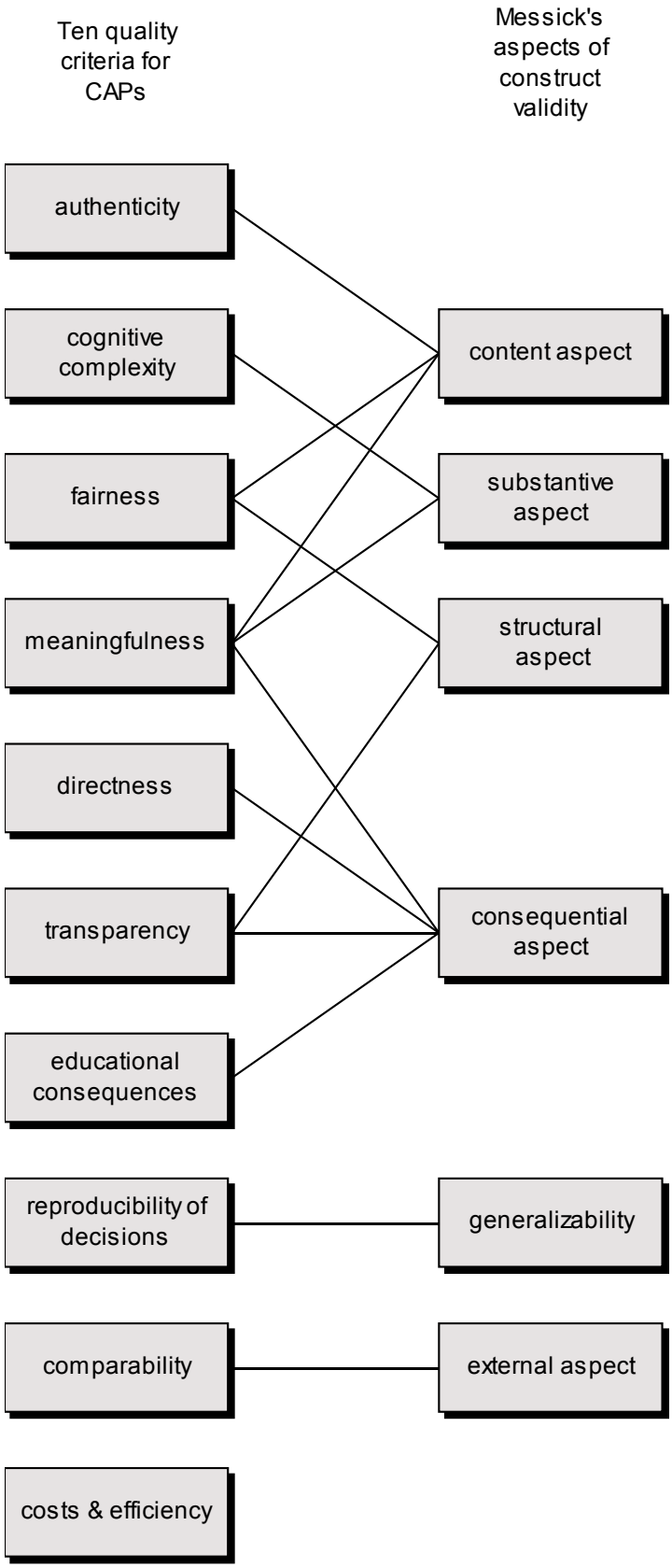
Criterion	Link to Messick's aspects	Analogies with Messick	Additions to Messick
Authenticity	Content	Inclusion of all knowledge, skills and attitudes to be measured	▪ Integrated assessment of knowledge, skills and attitudes ▪ Importance of work environment and social context
Cognitive complexity	Substantive	Measurement of thinking processes	▪ Assurance of cognitive complexity during the assessment itself
Fairness	Content	Inclusion of all knowledge, skills and attitudes to be measured	▪ Additional focus on recognizing individual differences between learners
	Structural	Match of criteria and weights with construct to be measured	▪ Focus on stakeholders' opinions in addition to technical elements of tests
Meaningfulness	Content	Meaningfulness of tasks and content	▪ Focus on stakeholders in assessment process (learners, assessors, work field) ▪ Focus on meaningfulness of feedback and criteria for learning process
	Substantive	Meaningfulness of thinking processes measured	
	Consequential	Link between meaningfulness and influence on learning made by other researchers (e.g., McDowell, 1995)	

| Criterion | Link to Messick's | Analogies with Messick | Additions to Messick |

	aspects		
Directness	Consequential	Focus on effect of different assessment forms on learning	▪ Additional focus on assessors or observers who have to interpret results
Transparency	Structural	Transparent link between scoring and construct structure	▪ Necessity of transparency to achieve other effects, for example effect on learning
	Consequential	Link between transparency and influence on learning made by other researchers (e.g., Gibbs, 1999)	
Educational consequences	Consequential	Focus on effects of assessment on teaching and learning	▪ CAP needs to have positive effect instead of just AN effect ▪ Assessment as part of the learning process and purposefully used to guide learning
Reproducibility	Generalizability	Increase in reproducibility implies increase in generalizability	▪ Focus on combining information sources instead of comparing different tests
Comparability	External	Prerequisite for generalisability or reproducibility	▪ Focus on conditions under which CAP takes place instead of different tests
Costs & Efficiency	-	-	▪ Additional focus on feasibility

Figure Caption

Figure 1. Qualitative Comparison of the ten-criterion framework for CAPs and Messick's (1994, 1995) construct validity framework


Ten quality criteria for CAPs
Messick's aspects of construct validity
authenticity
cognitive complexity
fairness
meaningfulness
directness
transparency
educational consequences
reproducibility of decisions
comparability
costs & efficiency
content aspect
substantive aspect
structural aspect
consequential aspect
generalizability
external aspect