1Running head: THE WHEEL OF COMPETENCY ASSESSMENT

This article was published as

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 Programmes. *Studies in Educational Evaluation, 32*, 153-170.

The Wheel of Competency Assessment: Presenting Quality Criteria for Competency Assessment Programs

Liesbeth K. J. Baartman*, Theo J. Bastiaens, and Paul A. Kirschner

Open University of the Netherlands

Cees P. M. van der Vleuten

Maastricht University, The Netherlands

---

* Correspondence concerning this article should be addressed to Liesbeth Baartman, Open University of the Netherlands, Educational Technology Expertise Center, P.O. Box 2960, 6401 DL, Heerlen, the Netherlands. E-mail: liesbeth.baartman@ou.nl

Abstract

Instruction and learning are increasingly based on competencies, causing a call for assessment methods to adequately determine competency acquisition. Because competency assessment is such a complex endeavor, one single assessment method seems not to be sufficient. This necessitates Competency Assessment Programs (CAPs) that combine different methods, ranging from classical tests to recently developed assessment methods. However, many of the quality criteria used for classical tests cannot be applied to CAPs, since they use a combination of different methods rather than just one. This article presents a framework of 10 quality criteria for CAPs. An expert focus group was used to validate this framework. The results confirm the framework (9 out of 10 criteria) and expand it with 3 additional criteria. Based on the results, an adapted and layered new framework is presented.

Keywords: evaluation criteria, alternative assessment, focus groups, vocational education

The Wheel of Competency Assessment: Presenting Quality Criteria for Competency Assessment Programs

Education is undergoing a global change from teacher-centred instruction for knowledge transfer, towards more learner-centred instruction and competency-based learning. This change has been set off by education's response to the changing labour market, which requires flexible, adaptive employees able to respond to a rapidly changing business environment, and who possess competencies instead of isolated knowledge and skills. These changes in instruction and learning necessitate the development of assessment methods to adequately determine the acquisition of those competencies.

The development of adequate assessment methods is of utmost importance because of the strong relationship that exists between learning and assessment. Alderson and Wall (1993) and Prodromou (1996) have described this as the "washback effect" or "backwash effect": what is assessed strongly influences what is learned. If assessment only measures factual knowledge, then learners will concentrate primarily on learning facts. Studies have shown that there is no greater impulse for learning than assessment (Frederiksen, 1984), with some authors even stating that any educational innovation will fail if there is no concomitant innovation of assessment (e.g., Cizek, 1997). Some authors (e.g., Biggs, 1996; Dochy, Moerkerke, & Martens, 1996; Tillema, Kessels and Meijers, 2000) see the linking of assessment to instruction as the cornerstone of success for the implementation of competency-based education. Biggs (1996, 1999) calls this constructive alignment, which does not prescribe a specific type of instruction, learning and assessment, but only prescribes that the three must be well-aligned. Such an alignment exists, for example, for traditional teaching aimed at knowledge transfer, rote learning and factual knowledge tests. However, since learning and instruction are increasingly

competency-based, this alignment is endangered because the development of adequate assessment methods appears to be lagging behind. If instruction and learning are based on acquiring competencies, then constructive alignment implies that assessment must also be competency-based.

A problem here is that the development of assessment methods to adequately assess the acquisition of competencies is hindered, because it is not clear what the requirements for these kinds of assessment are. Do traditional criteria for testing also apply to recently developed assessment methods or are other complementary or supplementary criteria needed? This article describes and evaluates a framework of quality criteria for competency assessments. First it argues that to adequately assess competencies, a combination of different methods in a Competency Assessment Programme (CAP) should be used. Second, it asserts that the well-known and widely used classical psychometric quality criteria of validity and reliability are not sufficiently suitable for evaluating the quality of CAPs. Based on a literature study a framework of quality criteria for CAPs is presented. This was evaluated in an international two-day expert focus group meeting. Based on the results of this meeting, an adapted and improved framework is presented.

From Methods to Programmes

Assessment of competencies is very complex, mainly due to the fact that a competency comprises a complex integration of knowledge, skills and attitudes (Van Merriënboer, Van der Klink, & Hendriks, 2002). Because assessing competencies is such a complex endeavour, it seems to be impossible to assess a competency using only one assessment method. The past ten years can be characterised by a transition from a testing culture to an assessment culture (Birenbaum, 1996, 2003), with the concomitant development of new assessment methods

promising a panacea for the determination of competencies. Although new forms of assessment have been developed, this article argues that classic tests should not be ignored and discarded beforehand, because any method may contribute to the complex job of determining whether a learner has acquired a competency. Van der Vleuten & Schuwirth (2005) argue that assessment should not be viewed as a psychometric problem to be solved for one single assessment method, bus as an instructional design problem that encompasses the entire range of assessment methods used within the curriculum. Therefore, this article argues for integrating different assessment methods into a Competency Assessment Program (CAP), in which newer forms of assessment can be used in combination with more classical methods.

Old and New Quality Criteria

Questions arise as to what constitutes a high-quality CAP and how this can be evaluated. The first question regarding quality criteria is whether CAPs have to be of equal quality as traditional forms of testing. The answer to this question is an unequivocal yes. Within competency-based education CAPs are used to make high-stakes decisions about learners and the importance of quality criteria for CAPs, thus, must not be underestimated. For classical tests, reliability and validity are generally used as measures of quality. For new forms of assessment, different and other quality criteria have been proposed (e.g., Guba & Lincoln, 1989; Linn, Baker, & Dunbar, 1991). Because the assessment methods included in a CAP as proposed in this article originate in both the testing culture and the assessment culture, quality criteria derived from both cultures might be needed. In the same way that classical testing methods should not be discarded for use in CAPs, measures of reliability or validity are not fundamentally wrong for CAPs, but they should be applied in a different way and be combined with other quality criteria that are especially important for competency assessment.

In the remainder of this section, an introduction into quality criteria used within the testing and assessment cultures is given, followed by a framework of ten quality criteria for CAPs in the next section.

Classical quality criteria: reliability and validity

Should reliability and validity thus be applied in the same way for CAPs as they are for classical tests? Benett (1993) and Kane (1992, 2004) argue that the fundamental principles of classical test theory may be applied to more qualitative assessments of competencies. Benett draws on classical test theory and examines how the different notions of reliability and validity may be applied in the context of assessments in the workplace. Although the idea of reliability is not fundamentally wrong, some problems exist with regard to the use of reliability for CAPs. Traditionally, reliability is defined in terms of the consistency of measurement over repeated occasions given fixed raters (Dunbar, Koretz, & Hoover, 1991) or the relationship of a single test item to the test as a whole. The first idea might be useful for psychological tests measuring unchangeable traits, but in education, changes in time are expected and even part of educational goals. The second idea might apply to long tests made up of small single items, but in competency assessment of whole task performance this does not. The traditional views of reliability, thus, cannot be applied to CAPs and competency-based education. Cronbach, Linn, Brennan, and Haertel (1997) describe some features of new forms of assessment that make traditional ways of analyzing measurement error inadequate, for example the fact that recent assessment are generally norm-referenced and the tasks used are open-ended and complex. We need to look for other measures to make sure the judgment process proceeds fairly and responsibly (Gipps, 1994; Moss, 1994; Uhlenbeck, 2002). Sluijsmans, Straetmans, & Van Merriënboer (submitted) and Benett (1993) argue that the traditional statistical procedures used

for objective tests to establish reliability are not appropriate for competency assessment or work-based learning. We should abandon the idea that assessment is an exact science in which a "true score" can be found (Gipps, 1994). Van der Vleuten and Schuwirth (2005) emphasize another problem in working with the traditional concept of reliability. They argue that reliability has often been confused with objectivity and standardization. In their view, reliability is not conditional on objectivity and standardization. Reliability can also be achieved with less standardized tests and more subjective judgments using, for example, human observers, as long as sampling is appropriate. Concluding, the idea of reliability is important for CAPs, but it needs to be defined and estimated in a different way than is done for classical tests.

With regard to validity, Kane (1992) suggests judging the validity of a test using qualitative data in an argument-based way: the interpretation and use of test scores is clarified and the plausibility of the arguments is evaluated. A well-known framework of quality criteria is that of construct validity described by Messick (1994, 1995), which describes six aspects of construct validity: content, substance, structure, consequences, externality, and generalizability. Messick's work originated from the classical notions of validity and reliability, but includes newer ideas of validity such as consequential validity. The problem with using validity for evaluating the quality of competency assessments is that many different definitions of validity are distinguished. 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). Kane (2004, p. 135) also describes validity in very

general terms: “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”. He further mentions a number of different types of validity: face validity, content validity, predictive validity, criterion-related validity and construct validity. Whereas Benett sees construct validity as just one of the types of validity, in Messick’s framework construct validity is used as the overarching concept of all types of validity. Messick’s inclusion of consequences in the unified concept of validity increased the scope of formulations of validity, which was also acknowledged by AERA et al, but the use of construct validity as a the whole of validity causes problems. Because so many different forms of assessment are used, many validity aspects are “hidden”, and heaped together, converting validity into one huge container concept. A second problem is that quality criteria concerning implementation of CAPs (e.g. costs and efficiency) are not included in classical frameworks. Construct validity concerns itself with the actions of the user of the assessment information, but is limited with respect to other stakeholders involved or to subsequent actions. To sum up, the confusion about validity not only causes practical implementation problems, but it is often not conceptually clear anymore what meaning of validity authors have in mind.

New forms of assessment, new and more quality criteria

The shift towards an assessment culture led to the use of more and other quality criteria than reliability and validity. This development started with Messick including consequential validity in his framework (1994, 1995). Linn et al. (1991) argue that it is appropriate to expand the idea of quality because of the different characteristics of new forms of assessment, such as the linking of assessments to the way in which learning occurs. A new framework is needed that

is more consistent with current theoretical understandings of the nature and potential uses of assessment. Linn et al. mention criteria such as consequences, transfer and generalizability, fairness, cognitive complexity, meaningfulness, content quality, content coverage and cost and efficiency. Uhlenbeck also mentions a number of new quality criteria: authenticity, content quality, domain coverage, comparability, impact and practicality. Other criteria are extrapolation, generalizability and accuracy (Sluijsmans et al., submitted). Including these newer criteria into a framework of quality criteria for CAPs may do justice to the unique character of competency assessment.

Framework of 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., Driessen, Van der Vleuten, Tartwijk, & Vermunt, 2005; Frederiksen & Collins, 1989; Gulikers, Bastiaens, & Kirschner, 2004; Hambleton, 1996; Linn et al., 1991; Kane, 1992, 2004; Schuwirth & Van der Vleuten, 2004; Sluijsmans et al., submitted; Uhlenbeck, 2002; Van der Vleuten & Schuwirth, 2005). Following Biggs (1996), the quality criteria most aligned with the characteristics of CAPs were chosen. The goal of the framework is to provide a clear definition of all criteria and avoid container concepts to enable a further operationalization of the criteria into an instrument in further studies. Criteria were described separately as much as possible. While the ideas of validity and reliability are incorporated, in the framework they are worked out in a way different from Messick (1994, 1995). The new ideas about the quality of assessments are also included. Together, they provide an integral framework of quality criteria for CAPs. This is not only necessary from a theoretical point of view, but also because judgements about the value and relative merits of new forms of assessments and CAPs will depend on the criteria used to evaluate them.

It is very important to note that not all methods included in a CAP must meet all criteria, but that the programme as a whole must. For a CAP as a whole, deficits in one criterion cannot be balanced out by high scores on another criterion. For the quality criterion authenticity, for example, Gulikers et al. state that objective knowledge tests may only be used for high-stake summative assessment if the purpose of the assessment is not to determine future functioning in the workplace. Because this article argues for a *program* of competency assessment, knowledge tests can be included in a complete CAP aimed at assessing competencies. Part of this CAP might be a very authentic performance assessment, while another part might be a test to determine underlying knowledge, preferably integrated with the performance assessment. The assessment programme as a whole is evaluated against the criteria, of which some methods may score high on some criteria and other methods on different criteria. Taking together all methods included in a CAP, all quality must be met. The remainder of this section defines the quality criteria proposed.

Authenticity relates to the degree of resemblance of a CAP to the future professional life. A CAP should assess those competencies needed in the future workplace (Gulikers et al., 2004). The authors distinguish five dimensions that can vary in authenticity: the assessment task, the physical context, the social context, the assessment result or form, and the assessment criteria.

Cognitive complexity resembles authenticity in the sense that it also relates to the processes applied in future professional life, but it focuses more directly on the fact that assessment tasks should also reflect the presence of higher cognitive skills (Hambleton, 1996; Linn et al., 1991). An assessment task, depending on the phase of education, should elicit the thinking processes used by experts to solve complex problems in their occupational field. In this

respect, Hambleton remarks that the use of performance assessments is no guarantee that higher cognitive skills are indeed being measured. This should, thus, always be thoroughly investigated.

Meaningfulness implies the fact that a CAP should have a significant value for both teachers and learners (Hambleton, 1996; Messick, 1994), to which the importance in the eyes of future employers could be added. A possible way to increase meaningfulness is to involve learners in the (development of the) assessment process. 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. An assessment might also become more valuable to learners when they themselves can determine when they are ready to take the assessment and can thus gain most profit from it.

Fairness specifies that a CAP should not show bias to certain groups of learners and reflect the knowledge, skills and attitudes of the competency at stake, excluding irrelevant variance (Hambleton, 1996; Linn et al., 1991). Possible causes of bias are improper adjustment to the educational level of the learners or tasks containing cultural aspects that not all learners are familiar with.

Transparency relates to whether a CAP is clear and understandable to all participants. Learners should know the scoring criteria, who the assessors are, and what the purpose of the assessment is. As a possible indication of the transparency of an assessment, Hambleton (1996) suggests to check whether learners can judge themselves and other learners as accurately as trained assessors.

Educational consequences is mentioned as a criterion for competency assessment by many authors (Dierick & Dochy, 2001; Linn et al., 1991; Messick, 1994; Schuwirth & Van der Vleuten, 2004) and pertains the effects a CAP has on learning and instruction. 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 activities accordingly (Linn et al., 1991). This criterion is also related to effects like washback (Alderson & Wall, 1993; Prodromou, 1995).

Directness considers the degree to which teachers or assessors can immediately interpret the assessment results, without translating them from theory into practice (Dierick et al., 2001). A theoretical test does not immediately show if a learner is competent in a job situation, whereas a performance assessment does. Some evidence can be found that direct methods of assessment predict success at work better than more indirect methods (Uhlenbeck, 2002). Note that this does not imply that more indirect methods such as knowledge tests cannot be included in a CAP.

Reproducibility of decisions relates to whether the decisions made on the basis of the results of a CAP are accurate and constant over time and assessors. This does not mean that a CAP must be objective (Schuwirth & Van der Vleuten, 2004; Van der Vleuten & Schuwirth, 2005). Using performance assessments, assessors subjectively judge the performance of learners. Important is that the decisions about the learner are made accurately and do not depend on the assessor or the specific assessment situation.

Comparability addresses the fact that a CAP should be conducted in a consistent and responsible way. The conditions under which the assessment is carried out should be, as much as possible, the same for all learners and scoring should occur in a consistent way, using the same criteria for all learners (Uhlenbeck, 2002). Possibilities to increase comparability include careful sampling across conditions and using a large sample across the content and situations of the competency at stake (Van der Vleuten & Schuwirth, 2005).

Costs and efficiency are especially important because CAPs are generally more complex than classical tests and more difficult to carry out (Linn et al., 1991; Uhlenbeck, 2002). This criterion relates to the time and resources needed to develop and carry out the CAP, compared to the benefits. Evidence needs to be found that the additional investments in time and resources are justified by the positive effects, such as improvements in learning and teaching (Hambleton, 1996).

To validate this framework an expert focus group was organised. The goal was to let the experts build a framework of quality criteria for CAPs themselves and compare this expert framework to the literature framework. This way, it was explored whether the quality criteria described in the framework adequately cover all important quality control issues, or whether some criteria are missing or redundant. The expert framework and the literature framework were then combined to achieve an integrated and complete framework of quality criteria for CAPs.

Method

Participants

Participants in the expert focus group were fifteen international experts (from Israel, the United States, England, Scotland, Germany, Norway, the Netherlands) on assessment and quality criteria for assessment. Twelve experts participated in a two-day workshop and three experts gave a written reaction. To guarantee a broad basis for the expert framework, the experts were selected based on their expertise within a broad field of assessment research and practices.

Materials

An electronic Group Support System (eGSS) was used to guide the discussions and guarantee individual and anonymous input from all experts. An eGSS is a computer-based information processing system designed to facilitate group processes. It allows collaborative and

individual activities such as brainstorming, idea generation, sorting, rating and clustering via computer communication. All participants are seated in front of a laptop connected to a network and a facilitator computer. All input from the individual laptops can be combined into the facilitator computer and shown on a screen in various ways. All input generated from the expert meeting was collected and saved through the eGSS. All discussions were video-taped.

Procedure

At the start of the expert meeting, the participants were asked to enter in the eGSS all quality criteria they thought to be important for CAPs. To prevent influencing the participants, only a very general introduction about quality of assessment was given and the 10-criterion framework was not presented yet. The facilitator was a professional eGSS-facilitator. This assured both seamless use of the system and impartiality of presentation. The participants were first given three minutes to enter as many criteria as they wanted, which were gathered by the system and presented as a list on the screen. They were then asked to review the criteria entered and add two more criteria not yet included. This resulted in a list of sixty quality criteria. This list was reviewed by the researchers to combine duplicate and comparable criteria. The resulting list (20 criteria) was discussed in a plenary session in order to achieve mutual understanding of the quality criteria and to generate a workable list of criteria. All criteria were explained and discussed, revealing that different words were often used for the same idea. The discussion resulted in a final expert framework of thirteen quality criteria. Next, the 10-criterion literature framework was presented, and the meaning of the ten quality criteria was explained and discussed. The participants then compared the expert framework and the 10-criterion literature framework in a matrix, in which the expert criteria were put in the rows, and the literature criteria were put in the columns. An extra column (Other) was included. A score from 1 to 10

was given in each cell for goodness of match. The results were again presented to the group and discussed.

Analysis

The results of the expert meeting were analysed using both quantitative and qualitative techniques. Quantitative data are the means for goodness of match given in the matrix. A mean goodness of match of six was chosen as an indication of a good match. This value was chosen as a minimum value for goodness for match because on a scale from 1 to 10, 6 generally indicates a passing grade. Due to the exploratory nature of this study, no statistical analyses were used. Qualitative data included the video-taped group discussions about the expert framework. All criteria in the framework were discussed during the meeting and the typed out tapes were used to distil the definition given by the experts to all quality criteria.

The literature framework was combined with the expert framework in such a way that the quality criteria were kept separated as much as possible in order to give clear definitions and prevent container concepts. For matches of 6 and higher the qualitative data were used to investigate whether the criteria in both frameworks had the same meaning. If this was the case, the name used in the literature framework was retained. If a criterion in the literature framework had more than one match with the expert framework, the criteria of the expert framework were included as separate criteria. The other way around, if two or more criteria in the literature framework had the same criterion in the expert framework as a match, the original literature criteria were retained. If a criterion in one of the frameworks had no match of six or higher, it was excluded from the framework.

Results

The quality criteria generated by the experts are shown in the first column of Table 1. The cells represent the means and SDs for goodness of match. Due to a technical failure, the answers of two experts were lost. All means of six and higher are underlined, denoting a match between a criterion entered by the experts and one of the criteria in the literature framework.

- Insert Table 1 about here -

As can be seen, all criteria in the literature framework except Directness have one or more counterparts in the expert framework. Not surprisingly, Transparency matches perfectly with Transparence (M = 10), and the two Fairnesses match perfectly (M = 10). Educational Consequences matches almost perfectly with Backwash (M = 9.6). The video-taped discussions show that the meaning of Backwash as discussed by the experts is comparable to Educational Consequences as described in the literature framework. One of the experts described Backwash as follows:

> There are a number of dimensions to it, which is (1) intended, unintended, (2) positive and negative. I also think it is quite unpredictable. It requires you to monitor, follow up and evaluate the evaluation constantly, because you think you are doing something which is in line with the instruction, and apparently something unpredictable takes place and it has the exact opposite effect. But the backwash effect or the consequential validity is not only related to the students, but to the wider context, the teachers, the organization itself, the curriculum developers.

Costs & Efficiency matches with Practicality / Usability (M = 8). It is described by the experts as: “All things that have something to do with organising an assessment” and concepts

like easy to use, feasibility, costs/resources, and organizable were included. Reproducibility of Decisions and Comparability both match moderately high with Reliability (M = 7.1 and 6.2 respectively). One of the experts emphasized not to confuse reliability with objectivity:

> Objectivity is always considered to be a very important part of reliability, and is often operationalized as the agreement between people assessing something. But objectivity and reliability are not the same, quite on the contrary (…) you can have high objective assessment, which is totally unreliable, and the other way around you can have subjective forms of assessment which can be quite reliable. (….) You could call it reproducibility, and document how much noise, or in other words, how accurate your assessment is.

Authenticity and Cognitive Complexity are combined into the well-know quality criterion Validity (M = 8 and 6.4 respectively). Validity indeed appears to be a container concept, which is shown by the relatively high goodness of match of validity on all criteria in the literature framework. Meaningfulness as defined in the literature framework falls into three different categories: Fitness for Purpose (M = 6.3), Acceptability (M = 6.3), and Fitness for Self-assessment (M = 6.3). Fitness for Purpose was described using arguments like "fitness for purpose in relation to the curriculum", "assessment should fall together with the content that is assessed" and "fit to context". Acceptability is described as: "The assessment has to be accepted by those in the profession (…) it has to do with the attitudes, the views. It's a policy question". Fitness for Self-assessment was described by one of the experts as: "Self-assessment is a potentially very dense concept. Assessment in relation to fairly explicit and understandable criteria in relation to how am I managing this very particular task in the best way". With Fitness for Self-assessment the experts also referred to the idea of self-regulated learning. Assessment

can play a role in the process towards more self-regulation by making clear what the criteria are, by showing weaknesses and by stimulating reflection on the learning process.

Of the literature framework proposed, Directness does not have a counterpart in the expert framework that has a goodness of match higher than six. Apparently, the experts did not come up spontaneously with a criterion comparable to Directness, or did not consider this criterion to be important. Finally, the category Other did not yield any marks higher than six. The criteria Robustness, Accessibility, Trust and Capable of Evaluation neither have a satisfyingly comparable criterion in the literature framework, nor were put in the category Other. Table 2 shows a short description of these criteria as mentioned by the experts during the discussion. Of these criteria, Trust is related to Fairness (M = 5.9) and Robustness is related to Educational Consequences (M = 5.0), but the marks are not high enough to justify a new criterion. Accessibility scores 4.2 on Fairness, and Capable of Evaluation scores 4.6 on Other. Both scores are not high enough, though, to justify the inclusion of the criteria into the framework.

- Insert Table 2 about here -

Conclusions

The goal of this study was to describe and evaluate a framework of quality criteria for CAPs, by comparing this literature framework to a framework generated by experts in a two-day expert focus group.

Implications for the Framework

The results of this comparison have a number of implications for the framework proposed. First, all ten criteria proposed, except for directness, were considered to be important for CAPs by experts. These nine criteria therefore are maintained in the framework. The criterion Directness is excluded from the framework because it did not have a counterpart in the expert

framework. The fact that the experts did not come up spontaneously with a criterion like Directness could be explained by the fact that theoretically Directness can be considered to be fairly similar to Authenticity when talking about motor skills. Assessment methods that measure motor skills in a more direct way generally also are more authentic, for example a direct performance assessment during an internship is both an authentic and a very direct measurement. When it comes to cognitive skills, Directness is included in the idea of Cognitive Complexity. To be able to measure thinking processes, one has to ask learners, for example, to think aloud and give a rationale for their actions. Being able to measure Cognitive Complexity thus already implies a more direct measurement. Moreover, the results show that the experts included Authenticity and Cognitive Complexity into their idea of Validity. Directness also scored fairly high on Validity, which could implicate that Directness is indeed comparable to Authenticity and Cognitive Complexity. Taken together, it was decided to exclude Directness from the framework with the notion that Directness is already being paid attention to within the criteria Authenticity and Cognitive Complexity.

Second, some criteria in the literature framework were combined in the expert framework. Both Authenticity and Cognitive Complexity had Validity as a counterpart in the expert framework. To prevent container concepts, which was one of the main reasons for creating the new framework, Authenticity and Cognitive Complexity are maintained in the framework as two separate validity criteria. Validity indeed appears to be a container concept including a little bit of almost all criteria in the literature framework and it thus not useful, justifying the creation of a new framework with clear and separate criteria. In the same way, Reproducibility of decisions and Comparability both matched with Reliability in the expert framework. The discussions made clear that Reliability as a concept is indeed confused with

objectivity, which should be prevented in a framework for CAPs. To enable clear definitions and specifications, Reproducibility and Comparability are maintained in the framework as separate criteria. We argue that these criteria together better represent the idea of reliability as it can be used for CAPs and do not carry the burden of years of discussion about the exact meaning and interpretation of them.

Third, Meaningfulness in the literature framework fell apart into three criteria in the expert framework: Fitness for Purpose, Fitness for Self-assessment and Acceptability. Apparently, Meaningfulness can be interpreted in different ways. Acceptability, for example, could be related to Meaningfulness because a CAP may be more easily accepted if it is meaningful or vice versa. Fitness for Purpose and Meaningfulness are probably related because a CAP may be perceived to be more meaningful if it is well connected to the purposes of the education provided. Fitness for self-assessment may be related to Meaningfulness because a CAP may be perceived as more meaningful if it stimulates self-regulated leaning, a quality expected of competent professionals. These three quality criteria indeed appear to be related to Meaningfulness as described in the literature framework, but they are not the same. Therefore, the three criteria Fitness for Purpose, Fitness for self-assessment and Acceptability are included in the framework as separate criteria.

Concluding, the framework proposed is adapted in the following way: the criterion Directness is excluded from the framework and three new criteria are added, namely Fitness for Purpose, Fitness for self-assessment and Acceptability. This results in a new framework of 12 quality criteria. Furthermore, the classical quality criteria Validity and Reliability indeed appear not to be fit for CAPs. These criteria are too broad in a competence context, as was seen by the

large amounts of discussion and disagreement between the experts about the meaning of both concepts and the experts' warning not to confuse reliability with objectivity.

Discussion

The Wheel of Competency Assessment

The original framework, which attempted to present an "orthogonal" view of quality criteria, has been modified to become a "wheel of competence" in which the interrelationships made visible during the focus-group meeting are also visible (see Figure 1). Note that the neighbourhoods of the different cells within the layers and between the layers are arbitrary and contain no specific information.

- Insert Figure 1 about here -

In this "Wheel of Competency Assessment", Fitness for Purpose is in the middle of the wheel and is the basis for the development of all CAPs. Fitness for Purpose was shown to be comparable to the idea of constructive alignment (Biggs, 1996, 1999), which prescribes that all CAPs must be aligned with the goal of the learning process (i.e., the acquisition of competencies), and with the instruction given. The next and inner layer of quality criteria consists of Comparability, Reproducibility of Decisions, Acceptability, and Transparency. These are the more basic quality criteria for CAPs, which are already more commonly used in practice for the evaluation of assessments. The outer layer of criteria consists of Fairness, Authenticity, Cognitive Complexity, Meaningfulness, and Fitness for Self-assessment. These criteria generally are newer and originate in the assessment culture. We expect them to be less commonly used in practice than the criteria in the inner layer. The criteria are represented in layers or circles to represent the idea that they are interrelated. In the wheel, the criteria in the inner layer tend to be prerequisite for the criteria in the outer layer. The criteria in the inner layer probably are also

addressed first when designing CAPs, followed by the criteria in the outer layer, which on their turn build on the inner layer. For example, a CAP cannot be fair without being comparable and reproducible, and must be transparent before it can be perceived as meaningful. The square around the wheel represents the broader educational space in which assessment is taking place and here are two (possibly conditional) criteria, Costs & Efficiency and Educational Consequences. Educational Consequences pertains the relation between the assessment and education in general. Assessment, especially summative assessment, can have far reaching consequences for the student. On the other hand, formative assessment affects both learner choices, motivation and curriculum development and revision. The entire CAP should be of high quality to ensure that positive effects on learning and education in general are attained, as was also argued by Biggs (1996) and Tillema et al. (2000). Future research on Educational Consequences is needed to answer the question whether all quality criteria are needed to attain positive effects on learning, or whether some criteria are more influential than others. Costs & Efficiency pertains to another conditional relationship between assessment and education in general. As a part of an educational system, time and money needs to be allocated to all parts of education, of which assessment is just one. A CAP can be correctly designed according to all criteria, but if it cannot be implemented and used because of prohibitively high costs or low efficiency, the development has been a waste of time.

Future Research

The framework presented and validated in this study forms the basis for designing effective CAPs. This study has a very exploratory nature and the results should be interpreted with some caution. Further and more quantitative research is needed to further validate the framework. We strongly believe, though, that the application of the criteria helps answer the

need for assessments suitable for the determination of competency acquisition. A follow-up study investigating teachers' opinions about the importance of the quality criteria is currently being carried out. As was already mentioned, further research is also needed into the exact relationships between the quality criteria. For practical use, the criteria need to be further operationalized into a more practically oriented instrument which helps educational institutes to evaluate the quality of their CAPs.

References

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

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

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

Biggs, J. (1999). Teaching for quality learning at university. Buckingham, UK: SRHE and Open University Press.

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, MA: 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 Academic Publishers.

Cizek, G. J. (1997). Learning, achievement, and assessment: Constructs at a crossroad. In G. D. Phye (Ed.), Handbook of Classroom assessment: Learning, achievement, and Adjustment (pp. 1-32). San Diego, CA: 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.

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., Moerkerke, G. & Martens, R. (1996). Integrating assessment, learning and instruction: asessment of domain-specific and domain-transcending prior knowledge and progress. Studies in Educational Evaluation, 22, 309-339.

Driessen, E. W., Van der Vleuten, C. P. M., Van Tartwijk, J., & Vermunt, J. (2005). The use of qualitative research criteria for portfolio assessment as an alternative to reliability evaluation: a case study. Medical Education, 39, 214-220.

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.

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.

Gipps, C. (1994). Beyond Testing: towards a Theory of Educational Assessment. London: RoutledgeFalmer.

Guba, E. A., & Lincoln, Y. S. (1989). Fourth Generation Evaluation. Newbury Park, CA: Sage Publications.

Gulikers, J. T. M., Bastiaens, T. J., & Kirschner, P. A. (2004). A five-dimensional framework for authentic assessment. Educational Technology Research & Design, 52, 67-87.

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.

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

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

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

McDowell, L. (1995). The impact of innovative assessment on student learning. Education and Training International, 32, 302-313.

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

Moss, P. M. (1994). Can there be validity without reliability? Educational Research, 23, 5-12.

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.

Sluijsmans, D., Straetmans, G., & Van Merriënboer, J. J. G. (submitted). A new approach in portfolio assessment: the Protocol Portfolio Scoring-Method.

Tillema, H. H., Kessels, J. W. M., & Meijers, F. (2000). Competencies 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. Unpublished doctoral dissertation, University of Leiden, ICLON Graduate School of Education, Leiden, 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.

Van Merriënboer, J. J. G., Van der Klink, M. R., & Hendriks, M. (2002). Competenties: van complicaties tot compromis. Over schuifjes en begrenzers. [Competencies: from complications to compromise] (Onderwijsraad report 20020382/598). Den Haag, the Netherlands: Onderwijsraad.

Acknowledgements

The authors want to thank all participants in the expert focus group for their useful comments and discussions.

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

Table 1

Means and SD of Scores Given by the Experts for the Goodness of Match Between the Criteria Entered in the eGSS and the Literature Framework

Expert criteria	Authenticity	Cognitive Complexity	Meaningfulness	Fairness	Transparency	Directness	Educational Consequences	Reproducibility of Decisions	Comparability	Costs & Efficiency	Other
Validity	**8 (1.76)**	**6.4 (3.27)**	5.7 (2.95)	4.1 (3.32)	3.3 (2.87)	4.7 (3.5)	4.9 (3.7)	4 (3.3)	3.1 (3.07)	1.3 (0.67)	1 (0)
Transparence	1.6 (1.9)	1.3 (0.95)	2.9 (3.11)	2.4 (1.9)	**10 (0)**	2.8 (3.01)	2.4 (2.55)	2.5 (2.17)	2.6 (2.22)	1.2 (0.42)	1 (0)
Reliability	2.2 (2.53)	2.5 (2.55)	3.1 (2.77)	2.9 (3.25)	2.2 (2.57)	3.2 (2.94)	2.2 (2.57)	**7.1 (3.51)**	**6.2 (3.79)**	1.4 (0.97)	1 (0)
Fairness	2.5 (2.55)	1.7 (1.94)	2.8 (2.94)	**10.0 (0)**	2.9 (2.77)	2.6 (2.22)	2.6 (2.76)	3.8 (3.71)	3.6 (3.44)	1.4 (0.97)	1 (0)
Practicality / Usability	2.7 (2.79)	1.9 (1.91)	4.0 (3.4)	2.0 (2.16)	3.0 (2.62)	3.1 (2.99)	2.6 (2.91)	2.3 (2.21)	2.1 (2.02)	**8.0 (2.71)**	1 (0)
Backwash	2.4 (2.55)	1.8 (1.69)	3.5 (2.99)	2.6 (2.8)	2.8 (2.44)	2.6 (2.41)	**9.6 (0.84)**	1.9 (1.52)	1.5 (1.08)	2.1 (2.28)	1 (0)
Fitness for purpose	5.9 (3.88)	5.7 (3.13)	**6.3 (3.47)**	3.1 (2.96)	2.7 (2.41)	3.4 (3.13)	5.5 (3.75)	2.2 (2.3)	2.1 (1.91)	2.5 (2.42)	1 (0)
Robustness	2.1 (2.02)	1.8 (1.69)	3.2 (2.94)	2.1 (1.79)	2.1 (2.33)	3.2 (3.36)	5.0 (3.77)	2.1 (2.28)	1.9 (2.33)	2.0 (2.21)	2.8 (3.79)
Acceptability	3.2 (3.16)	3.2 (3.16)	**6.3 (2.95)**	5.0 (3.2)	4.8 (3.82)	2.9 (2.51)	3.4 (3.27)	3.7 (3.3)	3.5 (2.37)	1.8 (1.62)	1 (0)
Accessibility	2.1 (2.23)	1.8 (1.69)	2.2 (1.99)	4.2 (3.79)	2.3 (2.75)	1.5 (1.08)	3.0 (2.62)	1.4 (0.97)	2.5 (2.17)	2.0 (2.16)	3.7 (4.35)
Fitness for self-assessment	3.3 (3.74)	4.0 (3.74)	**6.3 (3.59)**	1.6 (1.35)	3.7 (3.09)	2.0 (1.76)	4.2 (3.85)	2.2 (2.82)	2.7 (2.87)	1.3 (0.67)	1.9 (2.85)
Trust	3.6 (3.69)	1.8 (1.69)	4.3 (3.80	5.9 (3.31)	3.0 (2.79)	3.5 (2.84)	2.1 (2.6)	4.6 (2.67)	4.1 (2.85)	1.5 (1.2)	1.9 (2.85)
Capability of evaluation	3.1 (3.49)	1.8 (1.93)	2.4 (2.95)	2.1 (2.42)	3.8 (2.9)	1.9 (2.02)	3.4 (3.69)	1.5 (1.08)	2.6 (2.8)	2.1 (2.13)	4.6 (4.65)

Table 2

Description of Robustness, Accessibility, Trust and Capable of Evaluation Extracted From the Video-taped Discussions of the Expert Focus Group

Criterion	Description given by experts
Robustness	(…) It means that basically all systems you establish are subject to some dilution or corruption. (….) It is about how bad the system is going to be after three of four years after you invented it. (….) what will happen if this particular assessment, in this particular moment of time goes wrong, next week, next month. They ought to build in something like this [Robustness] on a time scale of month or years in terms of policy implications.
Accessibility	There are two things in it. One is that the language should be accessible. That's quite an issue, it can exclude people from the assessment, and has to do with backwash. And the other is physical accessibility.
Trust	(…) It has to do with fairness, with acceptability. It's an overarching idea that assessment is not about techniques, it's about creating trust in the assessment itself.
Capable of Evaluation	Once an assessment model has been evaluated, it should be capable of evaluation (…) many years are invested in development, and then five years later we ask how on earth do we tell whether this assessment is working, and we got no data to evaluate it.

Figure Caption

Figure 1. Adapted framework: the wheel of competency assessment.

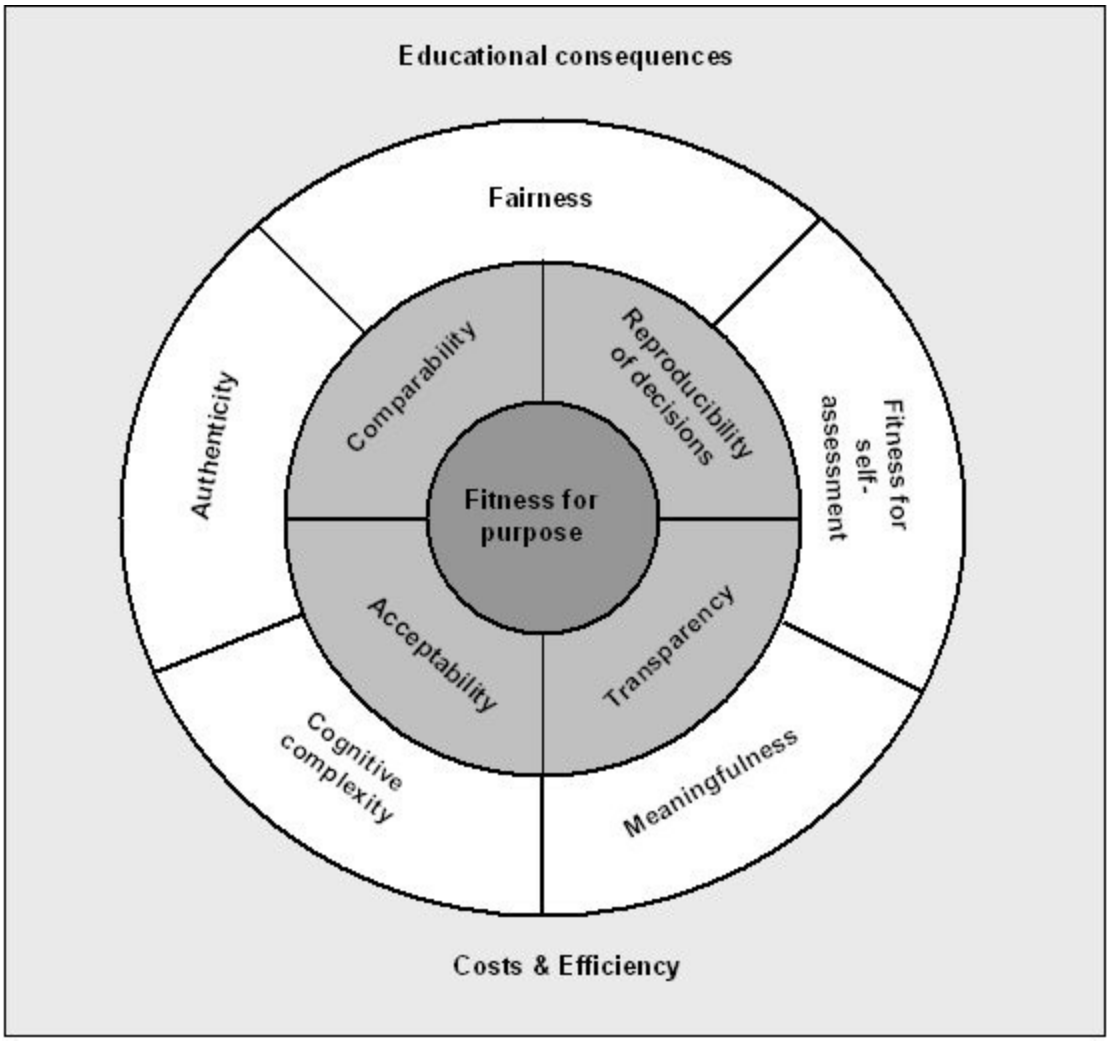
Educational consequences
Fairness
Fitness for self-assessment
Meaningfulness
Cognitive complexity
Authenticity
Comparability
Reproducibility of decisions
Fitness for purpose
Acceptability
Transparency
Costs & Efficiency

Figure 1.