‘Assessing the assessment’

Development and use of quality criteria for Competence Assessment Programmes

Liesbeth K. J. Baartman

Leden beoordelingscommissie:
Prof. dr. E. de Bruijn
Prof. dr. L. T. W. Schuwirth
Prof. dr. M. S. R. Segers
Prof. dr. K. M. Stokking
Dr. E. C. Roelofs

Netherlands Organisation for Scientific Research

The research reported in this dissertation was supported by the Netherlands Organisation for Scientific Research (project no. PROO 411-02-363).

ico

This research was carried out in the context of the Dutch Interuniversity Centre for Educational Research.

ISBN-978-90-393-4773-7
Printed by: Print Partners Ipskamp
Cover illustration: Anyldea, Zwolle

'Assessing the assessment'
Development and use of quality criteria for Competence Assessment Programmes

'De test getest'
Ontwikkeling en gebruik van kwaliteitscriteria voor Competentie Assessment Programma's
(met een samenvatting in het Nederlands)

Proefschrift

ter verkrijging van de graad van doctor
aan de Universiteit Utrecht
op gezag van de rector magnificus, prof. dr. J.C. Stoof,
ingevolge het besluit van het college voor promoties
in het openbaar te verdedigen
op donderdag 24 april 2008 des middags te 2.30 uur

door

Liesbeth Karina Joppa Baartman
geboren op 15 oktober 1979 te Culemborg

L.K.J. Baartman – 'Assessing the assessment' Development and use of quality criteria for Competence Assessment Programmes

Promotoren:	Prof. dr. P. A. Kirschner
	Prof. dr. C. P. M. van der Vleuten
Co-promotor:	Dr. F. J. Prins

Dit proefschrift werd (mede) mogelijk gemaakt met financiële steun van NWO.

Voorwoord

Veel mensen associëren het schrijven van een proefschrift met lange, eenzame computeruren op een klein kamertje. Oké, ik heb veel op mijn kamer zitten schrijven, analyseren, weer herzien, nogmaals proberen … en tja, wat dat betreft ben ik wel zo'n echte onderzoeker die dat ook leuk vindt. Maar minstens zo belangrijk zijn leuke collega's en een fijne werksfeer, en daarover heb ik zeker niet te klagen gehad.

Ik heb het geluk gehad bij twee universiteiten te mogen werken gedurende mijn aio-tijd. Ik ben begonnen bij de Open Universiteit, in het verre Heerlen, terwijl ik in Utrecht bleef wonen omdat we daar net een huis hadden gekocht. Overnachten deed ik twee keer per week in het NIVON-huis, vlak om de hoek bij de universiteit. Het was soms een heel gereis en geregel en veel mensen verklaarden me voor gek, maar ik heb er nooit moeite mee gehad "uit een rugzakje te leven". Bij de Open Universiteit heb ik het heel erg naar mijn zin gehad, en ik wil al mijn mede-aio's en collega's heel erg bedanken voor de kletspraatjes, de interesse in elkaar, en natuurlijk de Limburgse roddels. Met mede-aio's Ellen, Gemma, Amber, Wendy, Fleurie, Judith, Tamara, Desiree, Pieter, Marieke, PJ, Femke en Sandra heb ik vele lunchwandelingen gemaakt. Judith, wat mij betreft zetten we onze assessment-samenwerking voort. Amber, bedankt voor het hardlopen, samen eten en de grapjes over die "soms rare Limburgers". En Ellen, Gemma en Daniel, bedankt voor de Spaanse gezelligheid in Maastricht, Utrecht en Madrid!

De laatste anderhalf jaar van mijn aio-tijd heb ik bij de Universiteit Utrecht gewerkt. Tja, als je promotor én je co-promotor bij de Universiteit Utrecht gaan werken, en jij daar zelf ook woont … Gelukkig vond ik ook in Utrecht een hele leuke groep collega's, bij wie ik me al snel helemaal thuis voelde. Ook het geven van onderwijs was voor mij echt een mooie aanvulling op het onderzoekswerk. Ook hier weer de goede gewoonte van lunchwandelingen en - ik geef het toe - soms hopen op regen voor koffie bij Gutenberg. Harmen, Chris, Bert (overbuurmannen, of "de jongens" in de volksmond), Crina, Patrick, Agaath, Hendrien, Jeroen, Tim, en Sandy, bedankt voor de gezelligheid en de altijd openstaande deuren. Larike, heel leuk dat jij mijn ideeën weer verder uitwerkt in jouw research master. Marieke en Karel, ik kijk ernaar uit met jullie in het project Prove It te werken. En Elly, bedankt voor je enthousiasme bij de opzet van ons nieuwe onderzoeksproject. Ik ben heel blij dat ik nog een tijdje met jullie kan blijven werken!

Dan zijn er natuurlijk een paar mensen die heel belangrijk zijn in een promotietraject: de promotoren en co-promotoren. Paul, bedankt voor je nooit aflatende enthousiasme. Ik weet dat jouw deur altijd open staat. Cees, wat bewonder ik jouw vermogen om eerlijk en helder te zijn tegenover iedereen, en de manier waarop je je enthousiasme en inhoudelijke expertise weet in te zetten, en mensen tegelijk vrij te laten. Ik hoop daar zelf nog veel van te leren. Theo, jij hebt mij de eerste twee jaar op de Open Universiteit begeleid, en mij in de eerste zoekende periode binnen de paden weten te houden. Tot slot Frans, jij hebt de begeleiding van

Theo overgenomen toen hij naar Duitsland vertrok. Je hebt je supersnel weten in te werken in het onderzoek, en bent eigenlijk van alle markten thuis. Met jou kun je discussiëren, reflecteren, maar net zo gemakkelijk gezellig in een café zitten, en dat is een supercombinatie.
Tot slot wil ik hier de mensen van het Consortium Beroepsonderwijs bedanken voor de goede samenwerking, en de mogelijkheid bijna mijn gehele onderzoek bij jullie uit te voeren. Bartha Huijberts en Ellen Klatter, jullie hebben altijd opengestaan voor mijn ideeën, en ik word altijd weer enthousiast over competentiegericht onderwijs als ik zie hoe jullie bij het consortium daarmee bezig zijn. Alle opleidingen Laboratoriumtechniek van de ROC's die hebben meegedaan aan dit onderzoek, bedankt voor jullie durf in dit onderzoek te stappen en een zelfevaluatie uit te voeren van jullie assessments. Jullie hebben mij echt een kijkje in jullie keuken gegund.
Tot zover het werk. Mensen die mij goed kennen, weten dat ik graag op vakantie ga, en als een van de weinige aio's altijd mijn vakantiedagen heb opgemaakt ;-). Kitty, Jasper, Laura, Ellen, Paul, Linda, Vincent, Annet, Michiel, Floor, Giel, Marjolein, en de hele berg-groep, bedankt voor alle gezellige avonden thuis, op de klimmuur, in cafés en in berghutten, en de weekendjes en weken wandelen, klimmen, langlaufen en (toer)skiën. Ik ben er altijd helemaal uit als ik met jullie op pad ben, en jullie zijn altijd in voor nieuwe vakantieplannen. Papa en mama, niet alleen die liefde voor de bergen heb ik van jullie meegekregen. Jullie staan altijd voor ons klaar, onvoorwaardelijk, en ik ervaar het als iets ongelofelijk waardevols om zo'n basis te hebben meegekregen. Jantiene, wie heeft er nu zo'n superzusje voor wie ook nooit iets te gek is? En de allerbelangrijkste is en blijft natuurlijk Eric. Wat hebben wij het toch goed samen!

Liesbeth Baartman, februari 2008

Contents

1. General introduction

In many European countries, the ideas of competence-based education have got a firm foothold (Weigel, Mulder, & Collins, 2007), both at the level of policy making and at the level of educational practice. In the US, a similar movement towards what is called performance standards-based education can be observed (Valli & Rennert-Ariev, 2002). As a consequence of the shift towards competence-based education, assessment practices have to be changed as well, along with the ideas about what constitutes high-quality assessment. The subject of this thesis is the development, validation and practical use of a framework of quality criteria to evaluate the quality of assessment in competence-based education. Different from most research into assessment and assessment quality, this thesis focuses on assessment programmes, instead of on single methods. To elucidate this idea, the concept of Competence Assessment Programmes or CAPs is introduced. A CAP is a combination of both traditional and new assessment methods, which can have both formative and summative functions. This first chapter gives a general introduction into the studies conducted in this thesis. Three directions in assessment research, which form the basis for this thesis, are presented: competence-based education and the operationalisation of competence, the change from testing to assessment, and the change from psychometrics to edumetrics. The importance of thinking in terms of programmes of assessment is highlighted for each of the three directions. This is followed by the presentation of the main research questions. Next, an introduction is given into the context in which our studies were carried out - vocational education in the Netherlands - and the political and societal influences that currently are of great influence on assessment developments. This chapter concludes by presenting an overview of the studies presented in this thesis.

Directions in assessment research: The basis for this thesis

Theoretically, the studies described in this thesis can be embedded in at least three current assessment issues, namely (1) assessment content: competence-based education and the operationalisation of competence, (2) assessment methods: the change from 'testing' to 'assessment', and (3) assessment quality: the changed views on the quality of assessment, or from 'psychometrics' to 'edumetrics'. This general introduction describes our point of view on these issues, and all chapters further elaborate on them.

Assessment content: Competence-based education and the operationalisation of competence

Today's society and labour market pose new challenges to education and assessment. Modern societies are characterised by rapid technological changes and an exponential increase in available and accessible knowledge (Birenbaum, 2003; Tynjälä, 1999). Employees are not only expected to possess relevant domain-specific knowledge, but also to use this knowledge to solve increasingly complex problems, as well as to acquire new knowledge (Atkins, 1995; Tynjälä, 1999). Moreover, they are expected to work together in teams, to communicate with colleagues and clients, and to be critical thinkers, leading to an increasing focus on attitude as a necessary prerequisite for adequate functioning. Education, needing to provide learners with the necessary capabilities to function adequately in this changed environment, has adopted the ideas of competence-based education to support learners in meeting these new requirements.

Competence-based education and the concept of competence in itself can be operationalised in several different ways (Gonczi, 1994). First, a task-based or behaviourist approach arose in the 1960s and 1970s as a result of developments in society and the labour market and various publications on competence-based organisational training and teacher training in the US (Biemans, Nieuwenhuis, Poell, Mulder, & Wesselink, 2004). Task analysis approaches were used for example in the US, England and Australia to break jobs down into small behavioural subtasks resulting in skill-based instruction and training (Achtenhagen & Grubb, 2001). This approach was criticised for being reductionist in nature, for equating the task with the competence, for ignoring the influence of contextual and group factors, and for considering the aggregation of various atomised tasks to be sufficient to their integration – that is, that the whole was not greater than the sum of its parts (Gonczi, 1994).

The second approach views competences as general, stable and context-independent attributes underlying effective performance (Eraut, 1994). Problems with this approach are, for example, that it is questionable whether generic situation-independent competences actually exist (Gonczi, 1994) and the fact that novice-expert research has shown expertise to be highly domain-specific (e.g., Bereiter & Scardamalia, 1993; Greeno, 1989).

In response to the criticisms on the first two approaches, current operationalistions of competence move towards a third, more integrated approach, which tries to combine the idea of complex combinations of attributes (knowledge, skills, and attitudes) with the context in which these attributes are employed (Gonczi, 1994). The integrated approach parallels socio-constructivist theories of learning (e.g., Birenbaum, 2003; Tynjälä, 1999), which emphasise the idea of knowledge as something context-dependent, requiring meaningful learning activities and application in a realistic context (Brown, Collins, & Duguid, 1989). Not only is knowledge itself gaining in importance, but also how to cope with knowledge, how to transfer knowledge and how to use knowledge. Competence,

therefore, refers to the ability to integrate theoretical and practical knowledge and the capacity to learn from practical experiences (Atkins, 1995). Although the application of the constructivist philosophy to learning is criticised for not providing any evidence of improved learning outcomes or competence development (Kirschner, Sweller, & Clark, 2006), other studies have shown positive effects (e.g., Tynjälä, 1999). In our opinion, constructivism can provide a useful way of looking at learning processes that fits well within today's society with its rapidly increasing changes in knowledge in which employees must construct and reconstruct their expertise in a process of life-long learning.

This thesis adopts the integrated approach to competence and competence-based education, because such an approach acknowledges the context-dependent nature of competence, while acknowledging the importance of specifying competences and professional tasks at an appropriate level of generality (Hager, Gonczi, & Athanasou, 1994). As such, competence is defined here as the capacity to enact specific combinations of knowledge, skills, and attitudes in appropriate job contexts (Lizzio & Wilson, 2004). As the focus of this thesis is on assessment of competence, this means that the integrated approach specifies the content, or the 'what' of the assessment. To this end, we argue for a programme-approach to assessment because no single assessment method is sufficient to assess such complex wholes of knowledge, skills, and attitudes (e.g., Chester, 2003; Dierick & Dochy, 2001; Maclellan, 2004; Stiggins, 1991; Van der Vleuten & Schuwirth, 2005). A programme of different assessment methods could enable teachers and assessors to better capture the complexity of competence and as such generate a more valid picture of the student's development. This idea is further elaborated on in chapter 2.

Assessment methods: From testing to assessment to programmes

When education becomes increasingly competence-based, adequate assessment methods are needed to monitor and assess competence acquisition. Biggs (1996) presented the idea of constructive alignment, which prescribes that learning, instruction and assessment should be based on the same principles, in this case competence acquisition. From this perspective, it is inevitable that assessment methods are changed if approaches to learning and instruction change as described in the previous paragraph. Studies have shown that a strong relationship exists between learning and assessment (e.g., Frederiksen, 1984), implying that what and how is assessed strongly influences what is learned (e.g., Alderson & Wall, 1993; Gibbs, 1999; Myers & Meyers, 2007). Therefore, if assessment approaches are to be aligned with learning and instruction, they need to focus on the integration of knowledge, skills and attitudes. This means that more and different assessment methods should be used than 'just'traditional knowledge tests.

Table 1.1. Extremes of the testing culture and the assessment culture (partly derived from Segers, 2004, p. 9)

		Testing culture	**Assessment culture**
1	Content	Reproduction of knowledge	Multiple competences
2	Relation to learning process	Isolated	Integrated
3	Function	Summative	Formative
4	Context	Decontextualised	Contextualised (authentic)
5	Methods	Mainly (multiple choice) knowledge tests	Mix of different assessment methods
6	Responsibility	Teacher	Learner
7	Quality	Psychometrics	Edumetrics

The current views on learning and instruction have also changed the prevailing views with regard to the functions and methods of assessment. Some authors even speak of a paradigm shift or a transition from a testing culture to an assessment culture (e.g., Birenbaum, 1996; Birenbaum et al., 2006; Dierick & Dochy, 2001). A number of characteristics are often used to describe the differences between the testing culture and the assessment culture, in which the two are presented as extremes along a continuum (e.g., Segers, 2004). Although reality can rarely be adequately described as taking one of the extreme positions, it can be helpful to present the changes along a continuum to give an impression of the directions of change (see Table 1.1).

First, the content or the 'what' of assessment has changed, as was shown in the previous paragraph. While assessment in the testing culture mainly addressed lower level knowledge and skills, the assessment culture stresses the multidimensional nature of competence (Birenbaum, 1996). Second, while assessment was seen as isolated from the learning process in the testing culture, the assessment culture holds that learning and assessment should be interconnected (Wolf, Bixby, Glenn III, & Gardner, 1991). Assessment should not only focus on learning outcomes, but also involve the learning process leading to this outcome. The third change is related to the previous one and pertains to the changed function of assessment. In the testing culture assessment was mainly used as a summative measurement of what the student had learned at the end of the learning process. In the assessment culture, summative and formative functions are carefully balanced and assessment becomes part of a continuous cycle of assessment and feedback (Birenbaum, 2003). Fourth, assessment has changed from a decontextualised event to a relevant and interesting learning experience taking place in an authentic context, mirroring the importance of the context in the learning process in constructivist theories. Fifth, the assessment methods have changed. The most common assessment methods used in the testing culture are standardised tests, for example paper-and-pencil tests. The assessment culture added many different assessment methods, such as performance assessments and portfolios, which are used in different combinations over a prolonged period of time (Dierick & Dochy, 2001). The

sixth change relates to the responsibility of the assessment process, which has changed from being a sole teacher responsibility to a responsibility shared by the teacher and the learner, in which the learner gradually takes on responsibility for the learning and assessment process (Dochy, Segers, & Sluijsmans, 1999). Finally, the notion of what constitutes high-quality assessment has changed, a topic that is further addressed in the next paragraph.

As these examples show, testing and assessment are often presented as two extremes. During the transition from a testing to an assessment culture, many new assessment methods have been developed, which are sometimes referred to as 'alternative assessments'. This term might not have been very well chosen, as it implies that these assessments are replacements of, or alternatives to, existing methods. This thesis uses the term assessment to refer to all methods that can be used to determine and judge a learner's competence, whether originating in the testing culture or the assessment culture. Furthermore, it does not attempt to resolve the dispute between the testing culture and the assessment culture. Instead, it argues that (1) it is unwise to assume that new assessments methods are a panacea for the assessment of competence, and (2) assessment methods originating in the testing culture and the assessment culture should be viewed as having complementary rather than contradictory roles (Birenbaum, 1996; Maclellan, 2004). This is also again a reason to use programmes of assessment, in which traditional and new assessment methods are combined, and which can have both formative and summative functions. For example, the CAP of one the schools participating in our studies included written tests (both multiple choice and open questions), assessment of the products made during projects on which the students worked in groups, an assessment interview, and observations carried out by the teachers during classroom or practical work. Chapters 5 and 6 further elaborate on different assessment programmes and their characteristics and quality.

Assessment quality: From psychometrics to edumetrics

Assessment quality plays a key role in the transition towards assessment of competence. The development of assessment programmes to adequately assess competence acquisition could be supported if it is clear what the requirements for these kinds of assessment are. Do traditional criteria for testing - that is, validity and reliability - also apply to programmes of assessment which include both traditional and new methods of assessment, or are other complementary or supplementary criteria needed? As was done for assessment characteristics in the previous paragraph, this general introduction presents some changes in the prevailing views on assessment quality, or from the tradition of psychometrics to what is sometimes called edumetrics (e.g., Dierick & Dochy, 2001). The psychometric approach could be regarded as the description of what constitutes quality in the testing culture, while the edumetric approach is presented as an alternative to better account for the different characteristics of the assessment culture (Moss, 1994). Again, psychometric and edumetric approaches are often presented as two extremes, which are summarised in Table 1.2, and can be described as follows.

Table 1.2. Extremes of psychometric and edumetric approaches

		Psychometrics	**Edumetrics**
1	Object	Fixed traits	Competence development
2	Reference	Norm-referenced	Criterion / self- referenced
3	Reliability	Objectivity	Human observation
4	Measurement	Standardisation	Multiple measurements and generalisability
5	Function	Summative	Formative
6	Quality	Reliability is prerequisite for validity	Complementary quality criteria: e.g., meaningful, feedback, learning process

First, the psychometric approach stems from psychological measurements of fixed traits (e.g., intelligence or personality), based on which different characteristics of people could be distinguished. In the edumetric approach, on the other hand, the objects of measurement are not unchangeable personal traits, but the competence development of a learner, which is even expected to change over time (Wolf et al., 1991). Second, and related to the first difference, the psychometric approach sought to discriminate between individuals and used norm-referenced measurement in which different people are compared to each other. The edumetric approach uses a criterion-referenced approach, in which learners are not compared to each other, but to criteria that specify what should be learned, or a self-referenced approach in which learners are compared to their own past accomplishments (Martin, 1997; Sadler, 1987). Third, in psychometric traditions, reliability is generally achieved by standardisation and 'objectivity' (Birenbaum, 1996). Edumetric approaches acknowledge that competence assessment has to rely at least partly on human observation (Cronbach, Linn, Brennan, & Haertel, 1997), and shifted the focus from standardisation to the use of multiple measurements and generalisability across assessors and tasks. Fourth, edumetric approaches emphasise the importance of the formative function of assessment. This already implies the addition of other quality criteria than just reliability and validity, for example, that assessment should generate meaningful learning experiences, useful feedback, and stimulate the desired learning processes (Shepard, 2000).

What does the change from a psychometric to an edumetric approach mean for the quality of CAPs? Thinking in terms of programmes of assessment offers new possibilities to look at the quality of assessment, because the quality of the programme as a whole is evaluated. For example, the reliability pressure on formative assessments could be reduced and the resources freed up could be invested in the development of costly reliable and valid summative assessments (Knight, 2000). On the other hand, this thesis also argues that all assessments, including summative ones, have a 'formative potential' (Hickey, Zuiker, Taasoobshirazi, Schafer, & Michael, 2006) in steering students' learning processes. This means that learning-related quality criteria like meaningfulness and

educational consequences should apply to both formative and summative assessments in a programme. Chapters 2, 5 and 6 further elaborate on the issue of evaluating an assessment programme as a whole. With regard to quality criteria, we developed a framework of quality criteria for CAPs that originates in both psychometric and edumetric approaches. We thus incorporated psychometric ideas in our framework, but adapted them to make them more suitable for assessment programmes, and for competence-based education. In addition, edumetric ideas were added to do more justice to the nature of assessment in competence-based education (see chapter 2). The resulting framework of quality criteria for CAPs was validated by means of a teacher questionnaire and an expert focus group meeting. Chapters 3 and 4 further elaborate on this validation process.

Research questions

It is clear that assessment needs re-thinking in the direction of competence assessment, and that a programme-approach to assessment might be a valuable approach. It is not clear, however, what quality criteria are needed to evaluate the quality of these assessment programmes. This thesis focuses on this question, which can be divided into a number of sub questions, which are subsequently addressed in the studies described in this thesis:

1. What quality criteria are needed to evaluate the quality of assessment programmes in competence-based education?
2. How can these quality criteria be validated?
3. What is the utility of these quality criteria for practitioners?

This thesis reports on a series of studies working on these three research questions, and starting from the three current assessment issues described earlier. The main goal was to develop, validate and test the use of a framework of quality criteria that is suitable for evaluating the quality of CAPs in competence-based education. Theoretically, this thesis provides new ideas about how traditional psychometric criteria may be adapted for the use in competence-based education, and how they can be complemented with new quality criteria that do justice to new ideas about quality of assessment in competence-based education. Also, this thesis shows how quality criteria can be further operationalised into indicators, which adds to the transparency and understanding of what high-quality assessment programmes should actually look like. Finally, it shows how quantitative and qualitative data add to the evaluation of assessment quality, in which the value of qualitative approaches becomes especially clear. The practical relevance becomes clear in the third research question. One of the main arguments about assessment quality in this thesis is that quality is not only determined by the design of the assessment, but increasingly by its correct use in practice. Consequently, the usability of quality criteria becomes an important factor in the evaluation and development of high-quality assessment programmes.

Context of this thesis

Before presenting the overview of this thesis, this general introduction provides a description of the context in which our studies were carried out – vocational education in the Netherlands – and the political and societal factors that were of influence on our studies. As chapters 3, 4, 5 and 6 show, our studies are inextricably interwoven with the context in which they were carried out.

Vocational education in the Netherlands

In the Netherlands, there are a number of different forms of vocational education, each for different age groups and with different goals. Pupils can enter pre-vocational education at age 12. After leaving primary school, all pupils are required to enter secondary education where they choose between general secondary education (HAVO/VWO, age 12-17/18) and pre-vocational education (VMBO, age 12-16). Pre-vocational education serves as a preparation for vocational education (MBO, age 16-20), taken at a Regional Training Centre (in Dutch: Regionaal OpleidingsCentrum, ROC). When finishing vocational education, pupils choose either to enter higher professional education (HBO, age 17+) or to receive a vocational certification and enter the labour market. This thesis focuses on vocational education at age level 16-20.

Besides being divided into age levels, vocational education has a structure that corresponds to the different sectors in the economy, and training programmes are offered at four different levels. The four sectors are: technology, commerce/administration, services/health care, and agriculture. Level 1 prepares the student to carry out relatively simple executive tasks. Level 2 offers basic vocational training. This level is set by the government as the minimum qualification needed by all people to function adequately in the labour market. Level 3 prepares young people to become all round professional workers who carry out their tasks independently. The highest level 4 leads to middle-management jobs or specialist functions, and gives entry to higher professional education. Finally, there are two learning pathways: the day-release programme (BBL, BeroepsBegeleidende Leerweg) and the vocational training pathway (BOL, BeroepsOpleidende Leerweg). Day-release offers at least 60% of the training in a company, and a school week usually involves four days of practical training in a company and one day at school. Vocational training is a more theoretical pathway where the percentage of practical occupational training is between 20% and 60%. This programme usually involves a number of internships, whereas pupils spend most of their time at school (Cedefop, 2004; OECD, 2004). This thesis mainly focuses on the latter programmes, offered at levels 3 and 4 in the technical sector.

At the moment, the qualification structure upon which all educational programmes in vocational education are based, is under review. The reasons for change include the fact that there were too many qualifications (over 700) and that vocational education could not respond quickly enough to changes in the labour market (Dutch Eurydice Unit, 2006). Furthermore, innovations in the direction of

competence-based education were expected to better link educational programmes to job requirements and to close the gap that existed between the labour market and education (Biemans, et al., 2005; Tillema, Kessels, & Meijers, 2000). At the moment, Knowledge Centres for VET (Vocational Educational and Training) and Industry, which are organised per sector and involve representatives from social partners and vocational institutions, are developing new competence profiles. The new qualification structure should be more relevant to the labour market and society, easier to use, transparent and recognisable, and flexible and long-lasting (Dutch Eurydice Unit, 2006). It should give more freedom to vocational institutions to adopt innovative educational methods. The most important change is the development of competence profiles which specify the competences students should acquire to receive a certification. They are described at a higher aggregation level than the qualifications in the old national structure and are explicitly linked to practical application. Trials with the new competence profiles started in 2005. From 2010 on, Dutch vocational institutions are legally obliged to base their courses on the new competence profiles.

National consortium of vocational education

The studies described in this thesis were almost all carried out within a national consortium of vocational education (Stichting Consortium Beroepsonderwijs) that has the goal to stimulate educational innovation via a bottom-up strategy. Starting with problem-based approaches in 2000, the joint institutions are now working towards competence-based education. Each branch of study develops its own curriculum in author groups, in which teachers from all institutions of that branch are united. They collaboratively design practical work-based projects described in 'project books' to be carried out by groups of students in six to eight weeks. The projects are currently used by a number of institutions, but most institutions are still using the older problem-based (instead of competence-based) material developed earlier by the author groups, called 'unit books'. The main changes from the problem-based to the competence-based curricula, and from the unit books towards the project books, are an increased emphasis on the importance of assessment in the professional job context, the increased importance of attitudes as part of adequate functioning on the job, and the fact that the project books are less prescriptive and stimulate students to regulate their own learning (Klatter, 2006). With regard to assessment, the unit books and project books offer some suggestions for assessment methods (e.g., that it is a good idea to end each project with a presentation given by all students about their respective projects), but schools are free to design and implement their own assessment programme and modify the projects according to their specific needs.

For this thesis, the national consortium forms a common frame of reference for the vocational education teachers participating in the study described in chapter 3 (the pre-vocational teachers in this study were not part of the consortium), and for all schools participating in the studies described in chapters 5 and 6. This means that participants in all studies were consciously working towards competence-based

education, and although they differed in their experiences with and opinions of competence-based education, they can all be considered as holding relatively positive opinions towards innovations in education.

Assessment in vocational education: Political issues

With respect to assessment and specifically to its quality, many current political issues are of influence on vocational education. In 1996, the Adult Education and Vocational Training Act (in Dutch: WEB, Wet Educatie en Beroepsonderwijs) was implemented. One of the aims of the new act was to make the qualification structure for vocational and adult education more coherent, which led to the development of the national qualification structure described before. Another aim was to improve the quality of vocational education by giving more responsibilities to the institutions themselves, who were required to set up and maintain a quality assurance system, and carry out self-evaluations which formed the starting point for external evaluations carried out by the Inspectorate of Education. In 2001, the Dutch government expressed little trust in the quality of the examinations in vocational education (Deetman, 2001), which threatened the civil value of vocational diplomas. To improve the quality, a national Examination Quality Centre (EQC; in Dutch KwaliteitsCentrum Examinering or KCE) was established in 2002, which defined national standards for quality to which vocational institutions must conform. Until then, institutions for vocational education could have their examinations accredited by any of a number of different awarding bodies. After its establishment, the EQC became the only body to approve or sanction examinations. The role of the Inspectorate of Education was limited to supervising the EQC. As was done for assuring the quality of education as a whole, the institutions themselves were here too held responsible for developing and carrying out the examinations for the educational programmes they provide. This responsibility covers all aspects, including development, preparation, marking, and carrying out the examinations, and explication of examination regulations. Institutions are held responsible for quality assurance and for correcting any shortcomings that might be encountered. They are also accountable to the public. On a yearly basis, schools have to demonstrate that their examinations comply with the national quality standards, which is evaluated externally by the EQC during an audit in the institution. If the standards are met, the school receives its accreditation from the EQC, allowing it to examine and certify students. If an institution does not comply with the standards and does not show sufficient improvement in the next year, the Minister has the possibility of withdrawing its right to certify students. This measure has not yet been taken.

The EQC was established in 2002, and in 2004 the first national standards were put to use. These quality standards focused on: management and organisation of examinations, contracting out examinations, the examination process, examination products, and analysis and evaluation. An evaluation (KCE, 2005) showed that these first standards, which were based on the old qualification structure, were not suitable to evaluate assessments in more competence-based

educational programmes. Therefore the national standards and the audit procedures were redesigned during the school year 2005-2006. The new standards are formulated on a higher aggregation level and leave more freedom to individual institutions to develop their own examinations. Moreover, the new audit procedure was divided into two phases: in the first evaluation phase, the EQC judges the self-evaluation carried out by the institution and based on the outcomes, the elaborateness of the second verification phase is determined (proportional inspection). These new standards and procedure were put in use in 2006, at which time institutions could choose to be evaluated on the old or the new standards.

Still, the new external evaluation system did not function well. In April 2007, the Educational Council for Vocational Education (MBO Raad, 2007) reported that institutions experience a certain tension between the standards set by the Inspectorate – for the quality of education as a whole - and those set by the EQC – for assessment quality specifically. While instruction and assessment should ideally be integrated (the idea of constructive alignment as described by Biggs, 1996), the institutions have to account separately for the quality of education and the quality of examination, to two separate institutions using different standards and different procedures. Also, the institutions experienced difficulties distinguishing between formative and summative assessments, because they do not make such an explicit distinction themselves, whereas the EQC only evaluates the summative part. Moreover, an evaluation carried out by the Inspectorate in June 2007 (Inspectorate of Education, 2007) revealed the impracticality and often unreliability of the EQC audits. Therefore, in September 2007 the government decided to discharge the ECQ from its function of externally evaluating the quality of examinations in vocational education. From November 2007 on, this task is being carried out by the Inspectorate. Related to these developments, in October 2007 the government suggested developing national examinations for vocational education. This proposal is still being debated by the different political parties, but it would present a real shift away from the responsibilities given to individual schools.

As this thesis focuses on assessment quality in vocational education, these political developments were of considerable influence. Chapters 5 and 6 show how vocational institutions struggle with the responsibility to demonstrate the quality of their assessments, and how their innovations in assessment are influenced by the procedures and standards of the EQC. On the other hand, the studies also show how the EQC has increased the awareness of the importance of high-quality examinations.

Overview of the thesis

The first research question - what quality criteria are needed to evaluate the quality of assessment programmes in competence-based education - is addressed in chapter 2. This chapter reviews psychometric and edumetric approaches to assessment quality, and proposes a framework of ten quality criteria for Competence Assessment Programmes or CAPs. This framework is then compared to Messick's

(1994, 1995) well-known psychometric framework of construct validity. The aim of this comparison was to investigate whether quality criteria for competence assessment should be fundamentally different from traditional criteria such as Messick's and whether complementary criteria are needed to do more justice to the specific characteristics of assessment in competence-based education.

Chapter 3 and 4 address the second research question about the validation of the proposed quality criteria. The purpose of the study described in chapter 3 was to validate the framework of quality criteria from the practitioners' point of view. Using a questionnaire, teachers were asked about the importance of the quality criteria for their classroom practices. Also, the opinions of two different groups of teachers were compared. The first group worked in pre-vocational education, in a setting in which they were allowed freedom to develop their own assessments, independent of national examinations or requirements set by the government. The second group worked in vocational education, where institutions are struggling with the new standards set by the EQC and their new task of accounting for the quality of their assessments. Chapter 4 describes the validation and improvement of the framework by means of an expert focus group meeting. The goal of this meeting was to let the experts build a framework of quality criteria themselves through a group discussion guided by an electronic Group Support System. Then, the expert framework was compared to the literature framework developed in chapter 2. This way, it was explored whether the quality criteria adequately cover all important quality issues, or whether some criteria were missing or redundant.

The third research question about the utility of the framework of quality criteria for CAPs is addressed in chapters 5 and 6. The fifth chapter describes the development of a self-evaluation procedure, meant to assist schools to evaluate their own CAP based on the framework of quality criteria. For this self-evaluation, all quality criteria were further operationalised into indicators, more concrete aspects of a quality criterion in practice. Using a multiple case-study approach, it was explored whether schools are capable of evaluating their own assessments, and whether they can support their claims by means of examples or evidence (i.e., whether they could substantiate their claims). More specifically, the value of the use of a group interview to stimulate discussion and reflection, and the combination of different perspectives on assessment by including different functionaries in the group interview, was explored. Chapter 6 compares the CAPs of a more traditional and a more innovative school in order to explore how these two schools use the quality criteria to evaluate their assessments, if they use different approaches to assure CAP quality, and if the innovative school's CAP better complies with newer or edumetric quality criteria. It also offers some possible explanations for the differences found between the two schools.

Finally, the last chapter concludes this thesis by a general discussion. A summary is given of the main findings, and some critical remarks and challenges for further research are discussed. Also, practical implications of the studies are described, related to the societal and political influences described in this chapter.

References

Achtenhagen, F., & Grubb, N. W. (2001). Vocational and occupational education: Pedagogical complexity, institutional diversity. In V. Richardson (Ed.), *Handbook of research on teaching* (pp. 604-639). Washington, DC: American Educational Research Association.

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

Atkins, M. (1995). What should we be assessing. In P. Knight (Ed.), *Assessment for learning in higher education* (pp. 25-33). London: Kogan Page.

Bereiter, C., & Scardamalia, M. (1993). *Surpassing ourselves. An inquiry into the nature and implications of expertise*. Chicago: Open Court.

Biemans, H., Nieuwenhuis, L., Poell, R., Mulder, M., & Wesselink, R. (2004). Competence-based VET in the Netherlands: background and pitfalls. *Journal of Vocational Education and Training, 56*, 523-538.

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, 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.

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

Brown, J. S., Collins, A., & Duguid, P. (1989). Situated cognition and the culture of learning. *Educational Researcher, 18*, 32-42.

Cedefop (2004). *Vocational education and training in the Netherlands. Short description*. Cedefop Panorama Series. Luxembourg: Office for official publications of the European Communities.

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

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

Deetman, W. J. (2001). *Stuurgroep examens MBO, Advies examinering MBO* [Steering committee on examinations in vocational education, Advice on examinations in vocational education]. Advice to the minister of Education, Culture and Sciences, April 20, 2001, Den Haag, the Netherlands. Retrieved November 8, 2005 from http://www.minocw.nl/documenten/brief2k-2001-24055c.pdf.

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., Segers, M., & Sluijsmans, D. (1999). The use of self-, peer- and co-assessment in higher education. *Studies in Higher Education, 24*, 331-350.

Dutch Eurydice Unit (2006). *The education system in the Netherlands 2006*. The Hague, the Netherlands: Ministry of Education, Culture and Science.

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

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 students learn. In S. Brown & A. Glaser (Eds.), *Assessment matters in higher education* (pp. 41-53). Buckingham, UK: SRHE.

Gonczi, A. (1994). Competency based assessment in the professions in Australia. *Assessment in Education: Principles, Policy and Practice, 1*, 27-44.

Greeno, J. G. (1989). A perspective on thinking. *American Psychologist, 44*, 134-141.

Hager, P., Gonczi, A., & Athanasou, J. (1994). General issues about assessment of competence. *Assessment & Evaluation in Higher Education, 19*, 3-16.

Hickey, D. T., Zuiker, S. J., Taasoobshirazi, G., Schafer, N. J., & Michael, M. A. (2006). Balancing varied assessment functions to attain systemic validity: Three is the magic number. *Studies in Educational Evaluation, 32*, 180-201.

Inspectorate of Education (2007). *Enkele aspecten van de werkwijze van KCE. Rapport van een inspectieonderzoek* [Some aspects of the procedures used by the EQC. A report of the Inspectorate]. Utrecht, the Netherlands: Inspectorate of Education.

KCE (2005). *Standaarden en werkwijze KCE nieuwe stijl* [Standards and procedures Examination Quality Centre new style]. Amersfoort, the Netherlands: KCE.

Kirschner, P. A., Sweller, J., & Clark, R. E. (2006). Why minimal guidance during instruction does not work: An analysis of the failure of constructivist, discovery, problem-based, experiential and inquiry-based learning. *Educational Psychologist, 41*, 75-86.

Klatter, E. B. (2006). Competentiegerichte projectwijzers voor de lerende onderzoeker [Competence-based project books for the learning researcher]. *Develop*, 2, 24-34.

Knight, P. T. (2000). The value of a programme-wide approach to assessment. *Assessment and Evaluation in Higher Education, 25*, 237-251.

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.

Maclellan, E. (2004). Initial knowledge states about assessment: novice teachers' conceptualisations. *Teaching and Teacher Education, 20*, 523-535.

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.

MBO Raad (2007). *Competentiegericht beoordelen in het mbo* [Competence assessment in vocational education]. De Bilt, the Netherlands: MBO Raad.

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.

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

Myers, C. B., & Myers, S. M. (2007). Assessing assessment: The effects of two exam formats on course achievement and evaluation. *Innovative Higher Education, 31*, 227-236.

OECD (2004). *The role of national qualifications systems in promoting lifelong learning. Background report for the Netherlands*. Nijmegen, the Netherlands: Knowledge Centre for Vocational Training & Labour Market.

Sadler, R. (1987). Specifying and promulgating achievement standards. *Oxford Review of Education, 13*, 191-209.

Segers, M. (2004). Assessment en leren als twee-eenheid: onderzoek naar de impact van assessment op leren [the dyad of assessment and learning: a study of the impact of assessment on learning]. *Tijdschrift voor Hoger Onderwijs, 22*, 188-220.

Shepard, L. A. (2000). The role of assessment in a learning culture. *Educational Researcher, 29*, 4-14.

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

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 and Evaluation in Higher Education, 25*, 265-278.

Tynjälä, P. (1999). Towards expert knowledge? A comparison between a constructivist and a traditional learning environment in the university. *International Journal of Educational Research, 31*, 357-442.

Valli, L, & Rennert-Ariev, P. (2002). New standards and assessments? Curriculum transformation in teacher education. *Journal of Curriculum Studies, 34*, 201-225.

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

Weigel, T., Mulder, M., & Collins, K. (2007). The concept of competence in the development of vocational education and training in selected EU member states. *Journal of Vocational Education and Training, 59*, 51-64.

Wolf, D, Bixby, J., Glenn III, J., & Gardner, H. (1991). To use their minds well: Investigating new forms of student assessment. In G. Grant (Ed.), *Review of Research in Education, Vol. 17* (pp. 31-74). Washington, DC: American Educational Research Association.

2. Evaluating assessment quality in competence-based education: A qualitative comparison of two frameworks[1]

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 chapter argues for Competence 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 traditional and new forms of assessment, criteria used to evaluate CAP quality should be derived from both psychometric and edumetric approaches. 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.

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

Introduction

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 traditional 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 assessment of competence, 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 chapter is to provide a first step towards the solution of this

problem. A framework of ten quality criteria for competence assessment is compared to Messick's (1984, 1994, 1995) framework of construct validity, an older and well-know traditional 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 to do more justice to the specific characteristics of assessment in competence-based education.

This chapter starts with our definition of competence, assessment of competence and introduces the idea of Competence Assessment Programmes or CAPs. Then, the ten quality criteria for CAPs 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 is competence?

Before turning to matters related to the assessment of competence, 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 solve a problem adequately. 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 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 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, this thesis uses an integrated approach to competence (Gonczi, 1994), as was described in chapter 1. 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 thesis uses the definition given by Lizzio and Wilson who define competence as the capacity to enact specific combinations of knowledge, skills and attitudes in appropriate job contexts.

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). As was described in chapter 1, 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 paragraph shortly describes the development from the testing culture to the assessment culture and the reasons behind this transition. 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). 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 approaches to 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 integrated 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 thesis 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.

Competence 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 thesis considers both traditional tests and newer forms of assessment as necessary components of a Competence 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 warns that it would be unwise to assume that alternative assessments are the panacea for all assessment problems and Stiggins states that the challenge is to align the various assessment options we have to cover the broad array of achievement targets we value. Dierick and Dochy 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 (Chapter 5; Baartman, Prins, Kirschner, & Van der Vleuten, 2007), uses project-based education as a form of competence-based education. To assess students, this school uses 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 Competence Assessment Programmes

Now that the ideas of competence and Competence 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 research 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 thesis 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

The direct use of the traditional quality criteria of validity and reliability for the evaluation of CAPs causes a number of problems. Here, we elaborate on these problems and describe how the 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 to 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 generalisability 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, generalisability 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 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, & Cohen, 1996) and it is difficult to disentangle its many intertwined facets (Birenbaum, 1996). According to Crooks, Kane and Cohen (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 et al.). 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 to 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 chapter 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 links 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 chapter 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 et al., 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). 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. Therefore, the different quality criteria were kept separately as much as possible, and were not grouped under larger headings. The ideas of reliability and validity are incorporated in the framework, but are worked out in a different way. In addition, new quality criteria derived from the assessment culture are included. Together, they provide an integral framework of quality criteria for CAPs.

1. *Authenticity* as a quality criterion for assessment is generally described as the degree of resemblance of an assessment 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.
2. *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). To gain insight into the thinking processes applied by students, MacLellan (2004) suggests having them provide a rationale for their answer or action chosen.
3. *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).
4. *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).

5. *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. (1991) write that direct assessments of performance appear to have the potential of enhancing validity.
6. *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.
7. *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).
8. *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.
9. *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 traditional tests, this necessitates reliance on human judgment (see *reproducibility of decisions*). 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.

10. 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 2.1. 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 traditional and new quality criteria. As our framework comprises both traditional 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). Without wanting to deny the importance of Messick's work, the goal of this chapter 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.

A psychometric validity framework

Messick's (1984, 1994, 1995) framework of construct validity is depicted on the right-hand side of Figure 2.1. 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,

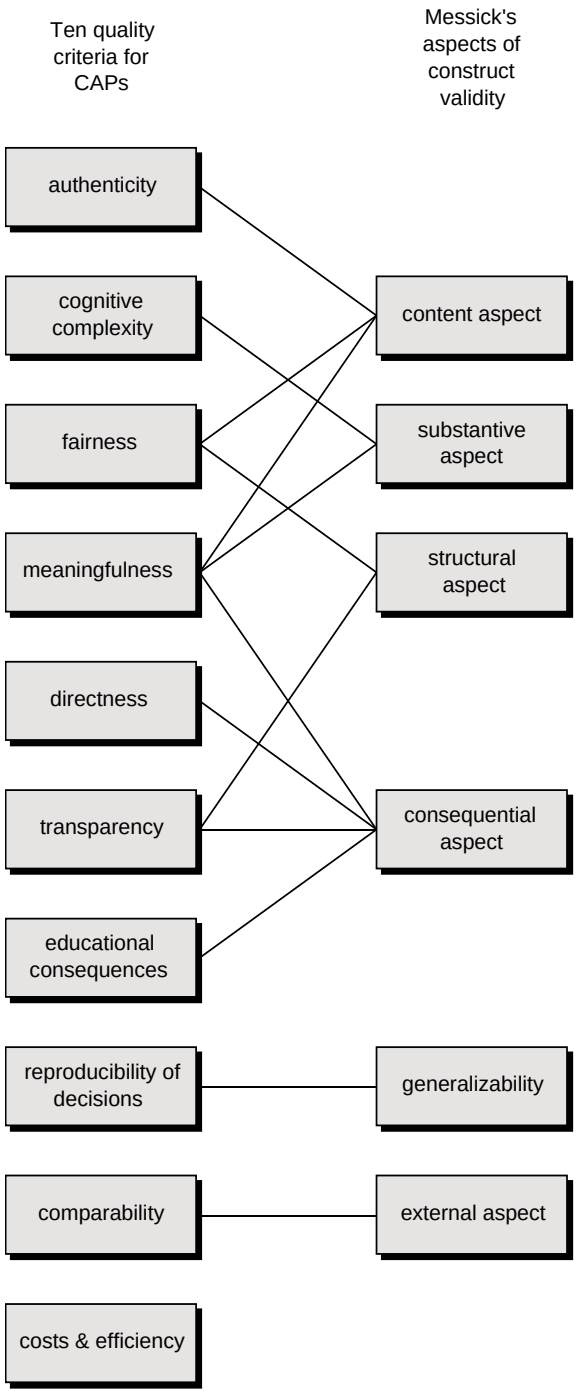

Figure 2.1. Qualitative comparison of the 10-criterion framework and Messick's construct validity framework

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, 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 (see also Figure 2.1). The differences pertain to 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 specifically important for assessing competences, an aim not included in psychometric frameworks. Table 2.1, starting on page 42, summarises the analogies and the differences between the ten quality criteria and Messick's framework.

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 (Gulikers et al., 2004). 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 the inclusion 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 to demonstrate 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 the content and structure of the assessment and the construct. *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 do 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, what needs to be added to Messick's framework are 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*

Table 2.1. Analogies and differences between the 10-criterion framework and Messick's (1994, 1995) aspects of construct validity

Criterion	**Link to aspects of construct validity**	**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)
	Substantive	Meaningfulness of thinking processes measured	Focus on meaningfulness of feedback and criteria for learning process
	Consequential	Link between meaningfulness and influence on learning made by other researchers (e.g., McDowell, 1995)	
Directness	Consequential	Focus on effect of different assessment forms on learning	Additional focus on assessors or observers who have to interpret results

Criterion	Link to Messick's aspects	Analogies with Messick	Additions to Messick
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	Generalisability	Increase in reproducibility implies increase in generalisability	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 and Efficiency	-	-	Additional focus on feasibility

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* is 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 makes 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 to 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. This criterion is very important for CAPs, because it can

never be successfully implemented if the costs are too high or if 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 chapter 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 approaches 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 traditional 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 traditional and new quality criteria (see also Figure 2.1). It also shows that, though related, they are operationalised in different ways and do not completely overlap (see Table 2.1). As predicted, our framework partly overlaps with Messick's, but also adds and elucidates quality aspects. First, as became apparent from Figure 2.1, 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 the 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 traditional and new quality criteria can be found, but when operationalised, the two traditions are clearly different.

Our second research question focused on whether the 10-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. Moreover, criteria like *meaningfulness* and *transparency* 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, teachers' and experts' opinions on quality criteria for CAPs are investigated in future studies (see Chapters 3 and 4). 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 institutions 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 instead of denying its importance. The ten criteria for CAPs are a first step in this direction. Together, they highlight relevant quality issues for CAPs in competence-based education.

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: American Educational Research Association.

Baartman, L. K. J., Prins, F. J., Kirschner, P. A., & Van der Vleuten, C. P. M. (2007). Determining the quality of Competences Assessment Programmes: A Self-Evaluation Procedure. *Studies in Educational Evaluation, 33*, 258-281.

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, 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.

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, England: 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, 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.

Crooks, T. J., Kane, M. T., & Cohen, A. S. (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 Press.

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, UK: 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, UK: SRHE.

Gonczi, A. (1994). Competency based assessment in the professions in Australia. *Assessment in Education: Principles, Policy and Practice, 1*, 27-44.

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 procedure for beginning teachers of English as a foreign language.* Unpublished 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 Education, 35*, 326-330.

3. Teachers' opinions on quality criteria for Competence Assessment Programmes[2]

Abstract

In the last decade, quality control policies towards Dutch vocational schools have changed. Schools must now demonstrate assessment quality to an Examination Quality Centre, a procedure that is usually carried out by the management without involvement of the teachers. This chapter argues that since teachers often design and carry out assessments, they must be involved in quality issues. This study therefore explores teachers' opinions on quality criteria for Competence Assessment Programmes, in order to validate a quality framework developed in earlier studies. Pre-vocational and vocational teachers (N = 211) responded to a questionnaire. Contrary to expectations, results show that teachers deem traditional and competence-based quality criteria equally important. Vocational teachers gave higher importance scores than pre-vocational teachers, possibly due to the pressure they experience to improve the quality of their assessments.

[2] This chapter is based on:
Baartman, L. K. J., Bastiaens, T. J., Kirschner, P. A., & Van der Vleuten, C. P. M. (2007). Teachers' opinions on quality criteria for Competency Assessment Programmes. *Teaching and Teacher Education, 23*, 857-867.

Introduction

There is a strong pressure on educational institutions and teachers to become more competence-based, so as to better meet the changing demands of the labour market. This has important consequences for student assessment because of the strong relationship that exists between instruction, learning and assessment. Assessment, learning and instruction should be aligned with each other (i.e. focus on the same learning outcomes). Also, assessment appears to strongly influence both how students learn and how teachers teach, causing both students and teachers to focus on what the assessment requires (e.g., Alderson & Wall, 1993; Biggs, 1999; Birenbaum, 2003; Frederiksen, 1984). A study focusing on how teachers connect instruction and assessment showed that teachers spend more than 35 % of their time on assessment and more than 10 % on assessment-driven instruction (Conca, Schechter, & Castle, 2004). A possible problem here is that whereas learning and instruction are increasingly competence-based, the development of adequate methods to assess those competences appears to be lagging behind. The past decade has seen a number of new assessment forms, such as performance assessment, authentic assessment and portfolio assessment (Gulikers, Bastiaens, & Kirschner, 2004; Hambleton, 1996; McDowell, 1995), each of which promises a panacea for the assessment of competence. But because competences are so difficult to assess, using one single assessment form seems not to be sufficient (Chester, 2003). Based on earlier work of the authors (Chapter 2; Baartman, Bastiaens, Kirschner, & Van der Vleuten, 2007), this chapter argues for a combination of different assessment forms - a Competence Assessment Programme (CAP) - which combines both traditional tests and recently developed assessment methods.

The use of CAPs in competence-based education seems promising, but teachers and educational institutions are struggling with how to determine the quality of the different assessment forms they use, both individually and in combination. Many teachers believe that they need strong measurement skills to construct assessments, and report a level of discomfort with the quality of their own assessments (Frey, Petersen, Edwards, Pedrotti, & Peyton, 2005). Two reasons can be given for this struggle. First, criteria such as validity and reliability, which have long been sufficient for traditional testing, are necessary but may not be sufficient for new assessment forms and combinations of these forms in CAPs (Moss, 1994; Taylor, 1994). Moreover, validity and reliability are defined and used in many different ways (Miller & Linn, 2000), which makes it difficult for teachers to effectively implement them in practice to evaluate their assessment methods. Maclellan (2004) showed that among novice teachers there was very little exemplification or elaboration of the concepts of validity and reliability and they did not connect issues of reliability and validity with different assessment methods. With the development of new assessment forms, concomitant, complementary or supplementary quality criteria have been proposed, such as the consequences, meaningfulness and cognitive complexity of the assessment (e.g., Linn, Baker, & Dunbar, 1991; Kane, 1992, 2004; Van der Vleuten & Schuwirth, 2005). As was argued in previous studies,

CAPs consist of combinations of both traditional tests and newly developed assessment forms, and therefore quality criteria from both traditional and new views on quality might be needed to evaluate their effectiveness (Chapter 2; Baartman, et al., 2007).

Second, apart from the fact that from a theoretical point of view it is unclear what quality criteria should be used for CAPs, educational institutions in the Netherlands are increasingly held responsible for demonstrating the quality of their assessments and are, therefore, looking for adequate criteria to evaluate their CAPs. An interesting case in this respect is vocational education. In the Netherlands, after leaving primary school, all pupils are required to enter secondary education. Here, they choose between general secondary education, which leads to entrance to a university or polytechnic, and pre-vocational education (age 12-16). Pre-vocational education serves as a preparation for vocational education, which can be taken at a range of levels (age 16-20). The Dutch vocational schools are comparable to the American vocational high schools. Almost half of the yearly cohort of Dutch pupils leaving primary school eventually enters some form of vocational education. When finishing vocational education, pupils choose either to enter higher professional education – comparable to vocational colleges or polytechnics – or to receive a vocational certification. In 2001, the Dutch government expressed little trust in the quality of the examinations in schools for vocational education (Deetman, 2001). To improve quality, the Examination Quality Centre (EQC) was established, which defined national standards for quality to which vocational schools must conform in order to retain their accreditation, put in use in 2004. In this vision[3], the schools are responsible for demonstrating that their examinations meet those standards. If the standards are met, the school receives its accreditation from the EQC, which allows examining and certifying their students. Without such accreditation, schools must enlist the services of another accredited institution. On top of this requirement, external monitoring has been increased to cover 100% of all examinations. The quality standards used by the EQC focus on: management and organisation of examination, contracting out examinations, examination process, examination products and accountability.

There is a problematic dichotomy here. First, the onus of proof of quality is shifted to schools, but schools seem not to be well-equipped – as an institution – to carry out this quality control. Vocational schools are struggling between the strict

[3] *Quality control policies in the Netherlands are developing in a direction opposite to countries like Great Britain and the US, where teachers' judgments are being replaced with external standardised tests. In the Netherlands, teachers' judgments are used and schools are free to design their assessments, providing they can demonstrate assessment quality to the national Examination Quality Centre. Looking at models of change (Bennis, Benne, & Chinn, 1969), the change is based on authority and the imposition of sanctions for failure; the power-coercive model of change. The difference seems to be that in the Netherlands authorities (i.e., the government and the EQC) pass on responsibility for demonstrating quality to schools, whereas in countries like Great Britain and the US responsibility is removed from schools, causing a feeling of loss of autonomy*

and often traditional standards set by the EQC and their wish to make education more competence-based (Onderwijsraad, 2006). Because teachers in vocational schools often design assessments, responsibility for quality control is also passed on to them. Second, since the teacher too is not especially qualified to carry out quality control, their individual credibility is threatened. The evaluation of assessment programmes is usually carried out by school management and external controlling bodies without involving the teachers working at the schools, although they are an important factor for achieving high quality CAPs. In the Netherlands, it is often the teacher who actually develops and carries out assessments and who has to play a role in the assurance of assessment quality. On the other hand, teachers have to work within an area of accountability and external control, which may threaten their credibility as teachers capable of their own assessment of student learning (Graham, 2005).

To assist both schools and their teachers, useful and usable quality criteria for assessments are needed. The goal of this study is to explore teachers' opinions on quality criteria for CAPs. In a previous study, a framework of ten quality criteria for CAPs was developed by means of a literature study (Chapter 2; Baartman et al., 2007). This framework is shortly described in the next section. The current study focuses on the validation of this framework by the actual users and developers of the assessment programmes, the teachers.

Ten-criterion framework for CAPs

Our framework of quality criteria is based on a synthesis of work by many authors (e.g., Driessen, Van der Vleuten, Tartwijk, & Vermunt, 2005; Gulikers, et al., 2004; Hambleton, 1996; Linn et al., 1991; Kane, 2004; Schuwirth & Van der Vleuten, 2004; Uhlenbeck, 2002; Van der Vleuten & Schuwirth, 2005). The goal of the framework is to provide a clear definition of all relevant criteria to enable their further operationalisation into an instrument for schools and teachers for evaluating CAPs. The framework comprises both criteria related to the traditional ideas of quality control such as *comparability*, *fairness*, *reproducibility of decisions* and *transparency*, and criteria that arose during the transition towards competence-based education such as *authenticity*, *cognitive complexity*, *costs and efficiency*, *directness*, *educational consequences*, and *meaningfulness*. Since CAPs consist of combinations of assessment methods, it is important to note that not all single methods included in a CAP must meet all criteria, but that the CAP as a whole must. For example, a non-authentic assessment form such as a written test for assessing knowledge about nurse-patient communication can be combined with a more authentic assessment form such as a performance assessment, in which the student really has to show his or her capabilities in communicating with patients. A CAP as a whole, on the other hand, has to comply with all quality criteria. For example, high scores on *authenticity* cannot offset deficits in *cognitive complexity*. Table 3.1 gives a short description of the ten criteria.

Table 3.1. Short Description of the Ten Quality Criteria for CAPs

Criterion	Short description
Authenticity	The degree of resemblance of a CAP to the future workplace, in terms of the assessment task, the physical and social context, and the assessment criteria
Cognitive complexity	A CAP should reflect the presence of the cognitive skills needed on the job, and should enable the judgement of thinking processes
Comparability	CAPs should be set up and carried out in a consistent way. The tasks, criteria, and working conditions should be consistent with respect to key features of interest, and scoring should occur in a consistent way.
Costs and efficiency	The feasibility of developing and carrying out the CAP, for both learners and assessors, and the time and resources needed, compared to the benefits
Directness	The degree to which teachers or assessors can immediately judge whether a student can function in a certain profession, without having to deduce or infer this.
Educational consequences	The degree to which the CAP yields positive effects on learning and instruction, and the degree to which negative effects are minimised
Fairness	CAPs should not show bias to certain groups of learners and reflect the knowledge, skills and attitudes at stake, excluding irrelevant variance
Meaningfulness	CAPs should have a significant value for all stakeholders involved. For learners, assessments should be a learning experience in themselves, and be useful for the learning process. For teachers and employers, the assessments should be meaningful in terms of the requirements of the future job
Reproducibility of decisions	The decisions made on the basis of the results of CAP should not depend on the assessor or the specific assessment situation. Therefore, multiple assessors, assessment tasks, and situations should be combined
Transparency	CAPs should be clear and understandable to all stakeholders. Learners and assessors should know the scoring criteria, and the purpose of the assessments. External controlling agencies should be able to get a clear picture of the way in which a CAP is developed and carried out.

The goal of this study is to explore the opinion of an important group of stakeholders in the assessment process, the teachers. First, the study investigates whether teachers consider quality criteria to be important for evaluating their assessment programmes and second, whether they deem some criteria more important than others. We expected teachers to deem traditional quality criteria more important than newer competence-based criteria, as teachers are often thought to be reluctant towards this change to competence-based education and assessment.

The distance between school managers and teachers seems to be increasing, causing teachers to only focus on their primary task of teaching, resulting in less commitment and awareness towards educational change (Onderwijsraad, 2006). The third goal of this study is to compare the views of teachers working in different types of education with different quality control policies. As described in the introduction, quality control policies in vocational education have changed dramatically in the Netherlands in the last half decade. This study compares the views of teachers working in vocational education to those of teachers working in pre-vocational education, the type of education leading towards vocational education. In 2001, a group of technical pre-vocational schools (called the 'ICT-route') got permission from the Dutch Ministry of Education to develop a new curriculum and assessment for technical pre-vocational education. They developed their own assessment programme, focusing on formative assessment and, working together with vocational schools, strived to permit a more fluid transition between pre-vocational and vocational education (Van der Sanden, Van Os, & Kok, 2003). Because assessment in these pre-vocational schools has a more formative and competence-based character, we expected teachers from these schools to deem newer quality criteria more important, whereas we expected teachers from vocational schools to deem traditional criteria more important. Using a questionnaire, teachers' opinions about the quality criteria were investigated and differences between vocational and pre-vocational education were studied.

Method

Participants

Enrolled in this study were 211 teachers, 40 of whom were working in pre-vocational education, and 171 in vocational education in the Netherlands. The teachers were working in different departments in schools throughout the Netherlands, including the personal and social services, health care, economics and technology sectors. The vocational schools were asked to participate through a national organisation of vocational schools. Eighteen school departments agreed to participate in this study. The teachers working at pre-vocational schools were contacted through the ICT-route school group. Of the 38 schools in the organisation, 34 agreed to participate in this study. Generally only one or two teachers per school participate in the ICT-route, resulting in 40 teachers participating in this study.

Materials

A questionnaire was developed based on the ten quality criteria of the theoretical study, in which the teachers were asked about the importance of the ten quality criteria for their assessments. The questions covered the theoretical definitions and descriptions of the quality criteria. As the quality criteria are fairly abstract concepts, the questions were formulated as examples of the quality criteria in practice. For example, in one of the *authenticity* questions, teachers were asked whether they deem it important to assess students in the workplace. Scales of four to eight questions were composed for each quality criterion. For the criterion *cognitive complexity* two subscales were developed, namely *thinking processes* and *thinking level*. The criterion *thinking processes* deals with the assessment of the way students think, make decisions, and provide a rationale for their decisions when performing a task. *Thinking level* pertains to the difficulty of the cognitive skills needed to solve the problems encountered on the job. At the end of the questionnaire, an open question enabled the teachers to give further comments on the quality of their assessments and their experiences with it. Table 3.2 presents the scales, the number of items in each scale and an example of an item of each scale. Answers on all questions were given on a 5-point Likert scale ranging from (1) not important at all to (5) very important. The last question was an open one in which the respondents were asked to express their opinion freely on quality criteria for assessments. In the instruction accompanying the questionnaire an explanation was given of a Competence Assessment Programme and the teachers were encouraged to give their personal opinion about the importance of the criteria: 'please give your personal opinion as a teacher, independent of current assessment practices and policy at your school. We would like to know what competence assessment should look like in your opinion'.

The questionnaire was pre-tested by a test panel of 10 teachers working in pre-vocational and vocational education. They filled out the questionnaire and commented on the readability of the questions and the (ir)relevance for vocational education. Based upon this pre-test unclear items were revised. In general, the examples of the quality criteria posed in the questions were considered to be understandable and relevant for teachers.

Before analysing the results of the questionnaire, the reliability of the criterion scales was determined. Table 3.2 also shows the Cronbach's Alpha scores of all scales, which were found to be moderately to highly reliable (range .59 to .82). To increase scale reliability, one question with a low item-total correlation value was removed from the *transparency* scale. To explore whether the criterion scales were uni-dimensional, a factor analysis was conducted on each scale. All scales proved to be uni-dimensional except for *cognitive complexity*. As was expected, this scale was composed of two distinct subscales. A factor analysis with Varimax rotation showed two factors, with Eigenvalue 3.57 and loadings ranging from .487 to .866 for *thinking level* and Eigenvalue 1.05 and loadings ranging from .661 to .737 for *thinking processes*. The first factor consists of all questions regarding *thinking level* and the second factor includes the questions about assessing *thinking processes*.

Table 3.2. Scales of the questionnaire filled out by the teachers

Scale	Cronbach's Alpha	Number of items	Illustration item
Transparency	.71	5	Students know and understand the assessment procedure
Authenticity	.69	5	Students are assessed in the workplace
Cognitive complexity *thinking level*	.78	4	The assessment task requires the thinking level needed in the future profession
Cognitive complexity *thinking processes*	.72	4	During the assessment students must justify and explain their decisions
Comparability	.82	4	The assessment tasks are equal for all students
Meaningfulness	.59	4	The school checks whether students deem the assessment task meaningful
Fairness	.73	6	The assessment method does not (dis)advantage certain groups of students
Costs and efficiency	.70	4	The time and money needed for carrying out an assessment are judged against the advantages of it
Educational consequences	.64	4	The school checks the effect of the assessment on student learning
Directness	.68	4	An assessor can directly observe whether the student is capable of functioning in a job
Reproducibility	.66	5	Multiple assessors are used for each student

Procedure

The questionnaires, which were in an electronic form, were distributed through a contact person at each school, usually the head of the department. The teachers received an e-mail from her or him with the request to fill out the electronic questionnaire on the Internet.

Analysis

The importance scores on all quality criteria were analysed by means of one-sample T-tests to investigate whether teachers consider the criteria to be important. The answers in the questionnaire were given on a 5-point scale, with 3 being neutral. When the scores given by the teachers were significantly higher than this neutral value, the criterion was regarded as being important in the eyes of the teachers.

Because many T-tests had to be used, Bonferroni corrections were applied. To test whether some criteria were deemed more important than others, an ANOVA was conducted with the judgement of the importance of the criteria as a within subjects-factor, since each teacher was asked to rate all criteria. In the same analysis, the level of education (pre-vocational or vocational education) was included as a between-subjects factor.

Results

The results are described in two sections. First, the perceived importance of the quality criteria is addressed, related to the questions whether teachers consider the quality criteria to be important and whether they deem some criteria more important than others. Second, the differences in importance of the criteria for the educational levels of pre-vocational and vocational education are described.

Perceived importance of the quality criteria

The mean importance scores of the quality criteria scales for the whole sample and for both types of education are shown separately in Table 3.3. On the one-sample T-tests all quality criteria were found to have scores significantly higher than 3 (*M* ranging from 3.88 to 4.50; $p < .001$ for all criteria) and were thus considered to be important. This was also the case for pre-vocational education (*M* ranging from 3.68 to 4.35, $p < .001$ for all criteria) and vocational education (*M* ranging from 3.96 to 4.54, $p < .001$) separately.

Table 3.3. Means and SD of criterion scales

	Overall		**Vocational**		**Pre-vocational**		**Diff.**
Criterion scale	***M***	***SD***	***M***	***SD***	***M***	***SD***	***M***
Transparency	4.50	0.59	4.54	0.57	4.35	0.65	.19
Authenticity	4.13	0.66	4.17	0.65	4.00	0.72	.17
Cognitive complexity *thinking level*	4.08	0.70	4.41	0.61	3.93	0.91	.48**
Cognitive complexity *thinking process*	4.08	0.72	4.07	0.70	4.10	0.79	-.03
Comparability	4.08	0.87	4.10	0.86	3.99	0.92	.11
Meaningfulness	4.05	0.67	4.04	0.66	4.10	0.69	-.06
Fairness	4.04	0.66	4.08	0.65	3.85	0.67	.23*
Costs and efficiency	4.04	0.80	4.13	0.77	3.68	0.84	.45*
Educational Consequences	4.03	0.71	4.05	0.71	3.95	0.71	.10
Directness	3.93	0.74	3.96	0.74	3.79	0.70	.17
Reproducibility	3.88	0.73	3.89	0.72	3.81	0.77	.08

* $p < .05$. ** $p < .01$.

The ANOVA yielded a significant main effect (Greenhouse-Geisser, $F(6.92, 1439.83) = 12.89$, $MSE = .38$, $\eta^2_p = .058$, $p < .001$), indicating differences in importance scores between the criteria. Post hoc tests (Bonferroni) were used to further investigate the differences between the criteria. Figure 3.1 shows the mean importance scores given by the teachers, together with the 95% confidence interval of the comparison between the different criteria. For easier comparison, the criteria in the figure have been ordered from most to least important.

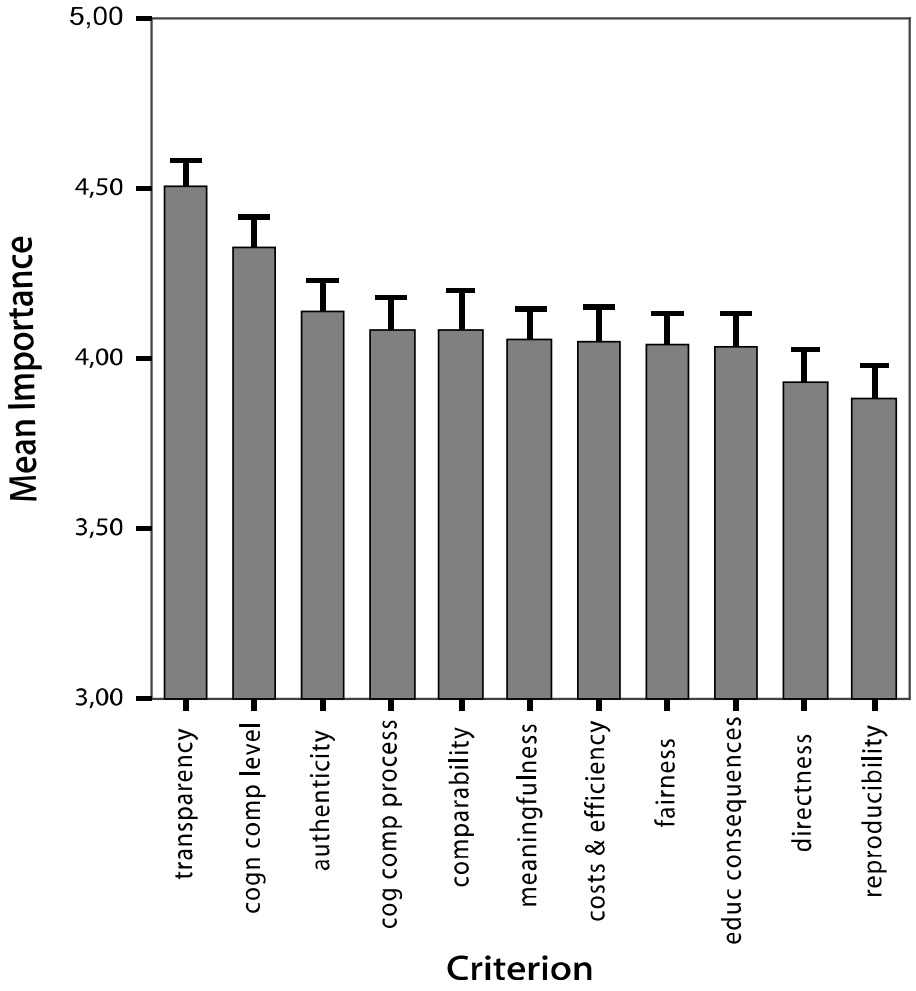

Figure 3.1. Overall mean importance scores with 95% confidence intervals

In general, the importance order seems to denote that quality criteria derived from traditional views (*comparability*, *fairness*, *reproducibility* and *transparency*) and newer criteria (*authenticity*, *cognitive complexity*, *costs and efficiency*, *directness*, *educational consequences* and *meaningfulness*) are considered to be equally important. This was confirmed by a paired samples T-tests comparing traditional and new criteria ($t(210) = 1.18$; $p = .238$). Regarding the importance of new quality criteria, derived from relatively new ideas about competence-based education, a division was noted between proponents and opponents of competence-based education. Some teachers elaborated on their opinions in the open question at the end of the questionnaire. A proponent of competence-based education wrote:

> My personal opinion is that each student enters the school with a number of competences. Our goal is to stimulate the student to develop these competences and to teach competences the student is less interested in. We have to assess what the student learns, and not what has to be learned. The personal interests of each student should guide the assessment. Each student should be able to get a certificate/diploma based on his or her competences. Assessment thus has to be very personalized.

On the other hand, an opponent of competence-based education expressed his opinion about the standards set for competence-based education. New standards have been formulated and schools have to prove their assessments cover these standards. Opponents of competence-based education state that the (factual) knowledge level of the student is decreasing because too much attention is paid to social and communication skills at the expense of knowledge:

> Right now we are focussing too much on communication, working in groups, etc … the level of education is decreasing … this is very bad, because until now companies were really satisfied about our education and I doubt whether this will remain so. Students at this level of vocational education will not lead discussions and give presentations in their jobs … we have to assess what they are going to do in their future jobs.

Comparing the criteria, *transparency*, which received the highest scores, was found to be significantly more important than all other quality criteria ($p < .001$). *Reproducibility* and *directness* received the lowest scores, and were found to be significantly less important than *transparency* ($p = .000$ for both), *cognitive complexity – thinking level* ($p < .001$ for both), *cognitive complexity – thinking processes* ($p = .005$ and $p = .017$ respectively), *authenticity* ($p = .006$ and $p = .01$ respectively) and *meaningfulness* ($p = .002$ and $p = .042$ respectively).

Differences between educational levels

The ANOVA also yielded a main interaction effect between the importance of the criteria and the educational level (Greenhouse-Geisser, $F(6.62, 1439.83) = 3.94$, $MSE = .38$, $\eta^2_p = .019$, $p < .001$). Independent T-tests were carried out to further investigate the differences in importance scores between pre-vocational and

vocational education. The differences between the two educational levels are depicted in Figure 3.2.

In general, the importance scores of teachers in both levels of education seem to show the same pattern. Overall, teachers in vocational education gave higher importance scores than teachers in pre-vocational education. In both types of education, *transparency* was found to be the most important quality criterion. The only two quality criteria which were judged as being more important in pre-vocational education are *meaningfulness* and *cognitive complexity-thinking processes*, but these differences were non-significant. Significant differences between the two levels of education were found for *cognitive complexity-thinking level* ($t(208) = -3.98$, $p < .05$), *fairness* ($t(208) = -2.00$, $p <$ 05) and *costs and efficiency* ($t(208) = -3.30$, $p < .05$), all of which were considered to be more important by teachers in vocational education

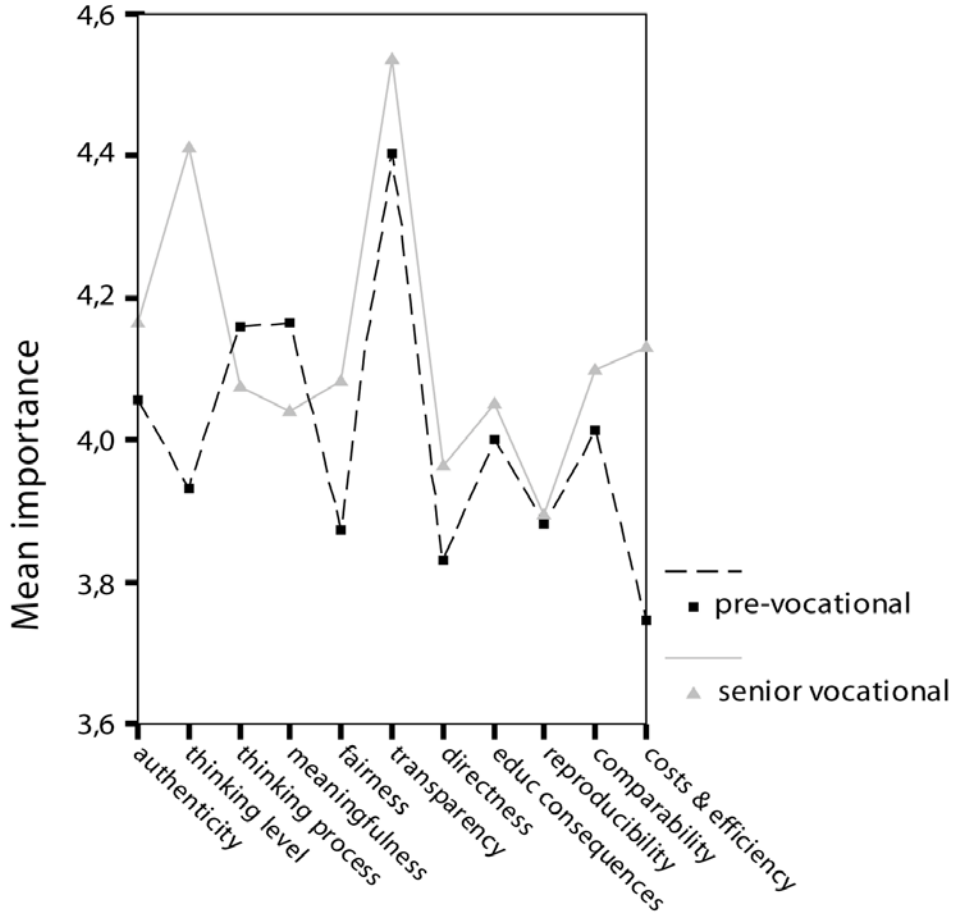

Figure 3.2. Mean importance scores in vocational and pre-vocational education

Conclusions and discussion

The goal of this study was to gain insight in teachers' opinions about the importance of quality criteria for CAPs, since teachers often develop and implement CAPs and play an important role in ensuring their quality.

The first research question focused on whether teachers considered the quality criteria to be important. The results show that this is indeed the case. As expected, all quality criteria were given high to very high scores on the importance scale, showing that teachers consider the quality criteria to be important for their assessments. On the other hand, a word of warning needs to be made, as the high scores and the small differences between the criteria might indicate socially desirable response behaviour or a high difficulty of the questionnaire. Also, these results do not mean they also actually carry out quality checks.

With regard to the second research question, the results show that teachers consider traditional criteria and newer criteria to be equally important. Again, the results need to be interpreted with some caution, as effect sizes were very low. Still, the results are interesting, as teachers are often thought to be reluctant towards adopting new assessment methods and criteria (Onderwijsraad, 2006). The discussion about whether or not it is necessary to complement the traditional views on quality control with new quality criteria has been going on for some time within the scientific field (e.g., Bachman, 2002; Moss, 1996; Webb, et al., 2003). This study adds a different point of view, that of teachers, to this discussion. The results seem to support the idea of combining both traditional and new views on quality control into an integral quality framework for CAPs. Also, the results show that, while all criteria were considered important, some criteria were deemed more important than others. *Transparency* scored very high, which may be due to the fact that in vocational education it was stressed by the Examination Quality Centre during their audits in the preceding years. The government's critics (Deetman, 2001) on the vocational examinations also addressed the lack of transparency and comparability between institutions. One of the main tasks of this new quality centre was to improve transparency. Whereas until 2001 they only evaluated 50 % of all summative examinations carried out at a school, they now check all of them, hereby expressing a clear wish to gain better insight in the assessment practices carried out at vocational schools. Increased transparency was needed for them to be able to achieve this insight. A second explanation of the high scores on *transparency*, which could also apply to pre-vocational education, is that, being in a transition period towards competence-based education, teachers experience many uncertainties in their work as a teacher, which increases their need for clarity about assessments. *Reproducibility* on the other hand scored relatively low compared to the other criteria. Assessing each student in different situations and the use of multiple assessors is often considered to be a possible solution to the reliability problems faced in competence-based education, but apparently teachers thought this relatively less important than other quality aspects. An explanation for the lower scores on *reproducibility* could be that teachers are not used to assessing students together with colleagues or other people and are afraid of loosing their autonomous position.

Being professional teachers, they possibly regard themselves as objective judges, hereby mistaking being a professional for automatically being objective. Another possibility is that the use of multiple assessors is just not a habit in vocational schools or teachers might think it is not feasible, being too costly and time-consuming.

The third research question pertained to the differences in opinions between pre-vocational and vocational education teachers. These results have to be interpreted with some caution, as group sizes between pre-vocational and vocational education were considerably different and the pre-vocational sample consisted of only 40 teachers. In general, teachers in vocational education gave higher importance scores than teachers in pre-vocational education. This may be because of the increased pressure to assure assessment quality that has been placed on vocational schools in the Netherlands by the new Examination Quality Centre. Vocational schools are not yet accustomed to being externally monitored and being responsible for demonstrating assessment quality themselves. The policy towards pre-vocational schools is more liberal. Moreover, pre-vocational education in general is not the end station of education. Consequently, assessment is not really used for certification, whereas in vocational education it is. In the Netherlands, there is a growing body of (public) opinion to put an end to summative assessments at the end of pre-vocational education. Instead, pre-vocational schools should and are often working together with schools to link up their curricula to permit a more fluid transition of students to vocational education.

Teachers in vocational education gave higher importance scores on *costs and efficiency*, *cognitive complexity – thinking level* and *fairness*. The higher scores on *costs and efficiency* indicate that teachers in these schools are more concerned that new assessment methods will be too expensive and too time-consuming. Until recently, these schools have had less opportunity to experiment with new assessment methods, which might explain this reluctance. The results also indicate that giving schools the opportunity and freedom to experiment with new assessment methods, as was done in our group of pre-vocational schools, could diminish reluctance towards these innovations. The fact that vocational education teachers judge *cognitive complexity – thinking level* to be more important can be explained by the fact that they are working at a higher level of education. As stated, pre-vocational education is not the end station for most students, while at the end of vocational education most students start working. At the end of pre-vocational education, the thinking level is still less important than it is at the end of vocational education, because this is when students need to be prepared to start working in a specific professional field with the accordingly required level of reasoning, or continue their education in institutions for higher vocation education, which also poses demands on thinking level. The higher *fairness* scores given by teachers in vocational education are not surprising since assessment in vocational education is generally meant for certification, whereas assessment in pre-vocational education is not. In the eyes of teachers, *fairness* is probably more important in summative than in formative assessment situations. Further research is needed here into the question whether the same quality criteria should apply for formative and summative assessments.

To conclude, this study presents teachers' opinions on a framework of quality criteria for CAPs, which includes both traditional and new views on assessment. As such, it adds a different point of view, that of teachers, to the scientific discussion about quality criteria for competence assessment. The framework provides an answer to the discussion about whether or not it is necessary to complement the traditional views on quality control with new quality criteria that may do more justice to the unique character of competence assessment. In this research, both views are combined into an integral framework and teachers seem to support this idea.

For practical purposes, this study provides a framework of quality criteria for the evaluation of existing CAPs and for the development of new CAPs suitable for competence-based education. It gives insight in teachers' opinions about the importance of the different criteria, which can help schools establish priorities in quality control issues. At the moment, the framework is more theoretically than practically oriented. In practice, it will probably be difficult to implement all quality criteria at the same time. Further research also needs to show whether the framework is applicable in all types and levels of education. This study was limited to vocational education, and criteria and priorities might be different in for example universities or online education.

After validation by experts in a next study (see chapter 4), an instrument will be developed using these criteria as a starting point, to support schools and teachers evaluating and improving the quality of their CAPs. Future studies (see chapters 5 and 6) will also investigate whether schools can really work with these kinds of criteria. It is important that all stakeholders in the assessment process accept such an instrument. This study guarantees teachers' opinions are taken into account.

References

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

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

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

Bennis, W. G., Benne, K. D., & Chinn, R. (1969). *The planning of change*. New York: Holt, Rinehart & Winston.

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

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.

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

Conca, L. M., Schechter, C. P., & Castle, S. (2004). Challenges teachers face as they work to connect assessment and instruction. *Teachers and Teaching: Theory and Practice, 10*, 59-75.

Deetman, W. J. (2001). *Stuurgroep examens MBO, Advies examinering MBO* [Steering committee on examinations in vocational education, Advice on examinations in vocational education]. Advice to the minister of Education, Culture and Sciences, April 20, 2001, Den Haag, the Netherlands. Retrieved November 8, 2005 from http://www.minocw.nl/documenten/brief2k-2001-24055c.pdf.

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.

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

Frey, B. B., Petersen, S., Edwards, L. M., Teramoto Pedrotti, J., & Peyton, V. (2005). Item-writing rules: Collective wisdom. *Teaching and Teacher Education, 21*, 357-364.

Graham, P. (2005). Classroom-based assessment: Changing knowledge and practice through preservice teacher education. *Teaching and Teacher Education, 21*, 607-621.

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

Maclellan, E. (2004). Initial knowledge states about assessment: novice teachers' conceptualisations. *Teaching and Teacher Education, 20*, 523-535.

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.

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.

Moss, P. A. (1996). Enlarging the dialogue in educational measurement: Voices from interpretive research traditions. *Educational Researcher, 25*, 20-28.

Onderwijsraad (2006). *Doortastend onderwijstoezicht. Aanbevelingen voor toekomstig toezicht op het onderwijs. Advies uitgebracht aan het Ministerie van OC&W*. [Vigorous inspection. Recommendations for future inspection of education. Advice to the Ministry of Education]. Den Haag, the Netherlands: Onderwijsraad.

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

Taylor, C. (1994). Assessment for measurement or standards: The peril and promise of large-scale assessment reform. *American Educational Research Journal, 31*, 231-262.

Uhlenbeck, A. M. (2002). *The development of an assessment procedure 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 Sanden, J. M. M., Van Os, M. J. M., & Kok, H. (2003). *Naar aantrekkelijk technisch vmbo. Resultaten van drie jaar herontwerp* [Towards attractive technical pre-vocational education. Results of three years of re-development]. Stichting Axis, Den Haag, the Netherlands: Opmeer Drukkerij B.V.

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

Webb, C., Endacott, R., Gray, M. A., Jasper, M. A., McMullan, M., & Scholes, J. (2003). Evaluating portfolio assessment systems: what are the appropriate criteria? *Nurse Education Today, 23*, 600-609.

4. The wheel of competence assessment: Presenting quality criteria for Competence Assessment Programmes[4]

Abstract

Instruction and learning are increasingly based on competences, causing a call for assessment methods to adequately determine competence acquisition. Because competence assessment is such a complex endeavour, one single assessment method seems not to be sufficient. This necessitates Competence Assessment Programmes (CAPs) that combine different methods, ranging from traditional tests to recently developed assessment methods. However, many of the quality criteria used for traditional tests cannot be applied to CAPs, since they use a combination of different methods rather than just one. This chapter 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.

[4] This chapter is based on:
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*, 153-177.

Introduction

Education is undergoing a global change from teacher-centred instruction for knowledge transfer, towards more learner-centred instruction and competence-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 competences 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 competences.

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 competence-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 competence-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 competence, then constructive alignment implies that assessment must also be competence-based.

A problem here is that the development of assessment methods to adequately assess the acquisition of competences 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?

From methods to programmes

Assessment of competence is very complex, mainly due to the fact that a competence comprises a complex integration of knowledge, skills and attitudes (e.g., Van Merriënboer, Van der Klink, & Hendriks, 2002). Because assessing competence is such a complex endeavour, it seems to be impossible to assess a competence 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 assessment of competence. Although new forms of assessment have been developed, this thesis 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 competence. Van der Vleuten & Schuwirth (2005) argue that assessment should not be viewed as a psychometric problem to be solved for one single assessment method, but as an instructional design problem that encompasses the entire range of assessment methods used within the curriculum. Therefore, this thesis argues for integrating different assessment methods into a Competence Assessment Programme (CAP), in which newer forms of assessment can be used in combination with more traditional 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 competence-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 traditional 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 thesis originate in both the testing culture and the assessment culture, quality criteria derived from both cultures might be needed. In the same way that traditional 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 competence assessment.

In the next sections, an introduction into quality criteria used within the testing culture and assessment culture is given, followed by a short description of the framework of ten quality criteria for CAPs.

Traditional quality criteria: reliability and validity

Should reliability and validity thus be applied in the same way for CAPs as they are for traditional tests? Benett (1993) and Kane (1992, 2004) argue that the fundamental principles of traditional test theory may be applied to more qualitative assessments of competence. Benett draws on traditional 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 competence assessment of whole task performance this does not. The traditional views of reliability, thus, cannot be applied to CAPs and competence-based education. Cronbach, Linn, Brennan, and Haertel (1997) describe some features of new forms of assessment that make traditional ways of analysing 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 (in press) and Benett (1993) argue that the traditional statistical procedures used for objective tests to establish reliability are not appropriate for competence 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 standardisation. In their view, reliability is not conditional on objectivity and standardisation. Reliability can also be achieved with less standardised 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 traditional 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 generalisability. Messick's work originated from the traditional 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 competence 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 of 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., (1999) 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 traditional 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 generalisability, 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. Sluijsmans et al. (in press) use the criteria extrapolation, generalisability and accuracy, which are again partly overlapping with the criteria used by Linn et al. and Uhlenbeck et al. Including these newer criteria into a framework of quality criteria for CAPs may do justice to the unique character of competence assessment.

Framework of quality for CAPs

The framework of quality criteria presented in this chapter 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., in press; Uhlenbeck, 2002; Van der Vleuten & Schuwirth, 2005). A more elaborate description of the framework is given in Chapter 2 and Baartman, Bastiaens, Kirschner, & Van der Vleuten (2007). The goal of the framework is to provide a clear definition of all criteria and avoid container concepts to enable a further operationalisation 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), as was shown in chapter 2. 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. (2004) 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 research argues for a programme of competence assessment, knowledge tests can be included in a complete CAP aimed at assessing competence. 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 criteria must be met. The remainder of this section defines the quality criteria proposed.

1. *Authenticity* relates to the degree of resemblance of a CAP to the future professional life. A CAP should assess those competences 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.
2. *Cognitive complexity* resembles authenticity in the sense that it also relates to the future professional life, but it focuses more directly on the fact that assessment tasks should also reflect the presence of the cognitive skills needed (Hambleton, 1996; Linn et al., 1991). An assessment task, depending on the phase of education, should elicit the thinking processes used by practitioners 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.
3. *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.

4. *Fairness* specifies that a CAP should not show bias to certain groups of learners and reflect the knowledge, skills and attitudes of the competence 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.
5. *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.
6. *Educational consequences* is mentioned as a criterion for competence 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).
7. *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.
8. *Reproducibility of decisions* relates to the fact that the decisions made on the basis of the results of a CAP should not depend on the assessor or specific assessment circumstances. This does not mean that a CAP must be objective (Schuwirth & Van der Vleuten, 2004; Van der Vleuten & Schuwirth, 2005). In many new assessments, assessors subjectively judge the performance of learners. Therefore, multiple assessors, assessment tasks and situations should be combined.
9. *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, comparable for all learners and

scoring should occur in a consistent way, using the same criteria for all learners (Uhlenbeck, 2002).

10. *Costs and efficiency* are especially important because CAPs are generally more complex than traditional 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. Vice versa, 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 4.1, on page 79. 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. 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 organisation itself, and the curriculum developers.

Costs and 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 'organisable' 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 operationalised 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',

Table 4.1. Means and SD of scores for the goodness of match between the expert criteria and the literature framework

Expert criteria	Authenticity	Cognitive complexity	Meaningfulness	Fairness	Transparency	Directness	Educational consequences	Reproducibility of decisions	Comparability	Costs and 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)

'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 4.2 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

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

Table 4.2. Description of robustness, accessibility, trust and capability 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 months 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.
Capability 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.

example, to think aloud or 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, 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.

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 thus 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 traditional 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 competence 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 4.1). Note that the neighbourhoods of the different cells within the layers and between the layers are arbitrary and contain no specific information.

In this wheel of competence 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 competence), 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 we expect to be already more commonly used in practice for the evaluation of

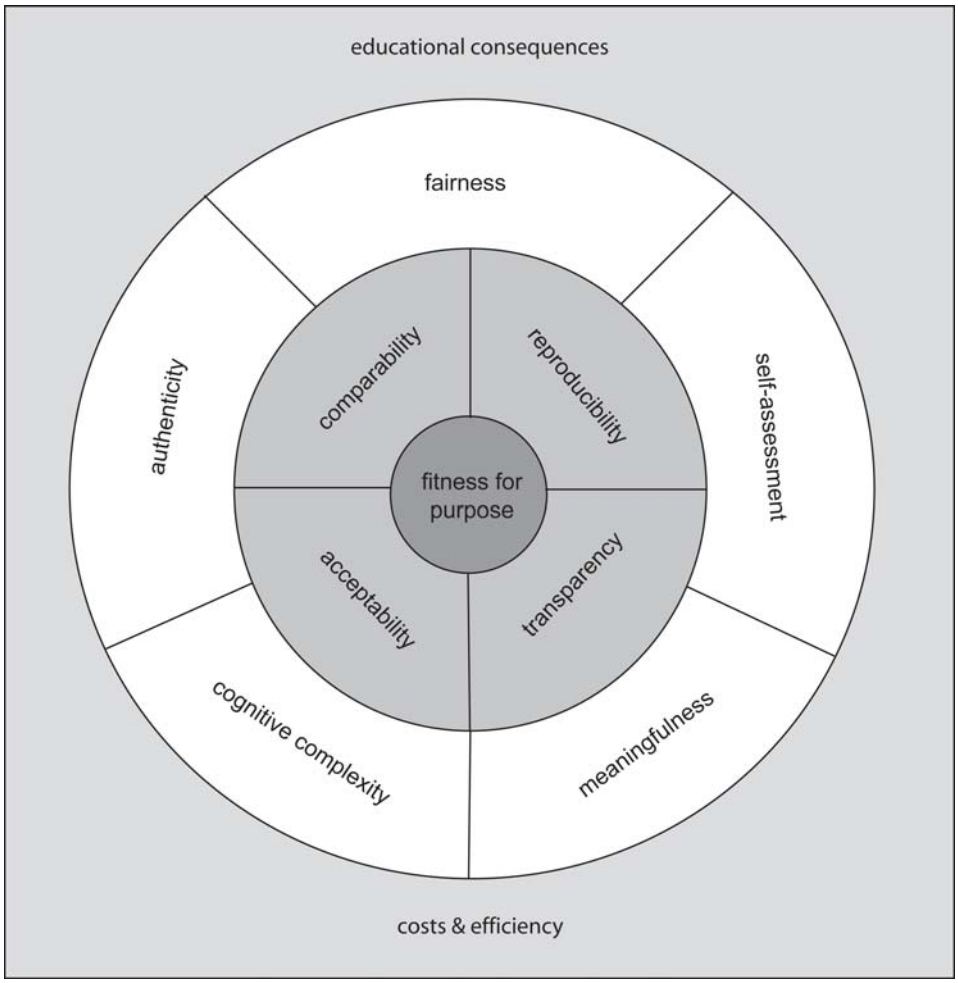

Figure 4.1. The wheel of competence assessment

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 takes place and here are two (possibly conditional) criteria, c*osts and 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 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 and 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 competence acquisition. To further study the quality criteria, a parallel study investigating teachers' opinions about the importance of the quality criteria has been carried out, of which the results were described in Chapter 3. 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 operationalised into a more practically oriented instrument which helps educational institutions to evaluate the quality of their CAPs.

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: American Educational Research Association.

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

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: assessment 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: Routledge Falmer.

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., Baker, 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. (in press). *A new approach in portfolio assessment: the Protocol Portfolio Scoring-Method*. Journal of Vocational Education and Training.

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 procedure 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.

5. Determining the quality of Competence Assessment Programmes: A self-evaluation procedure[5]

Abstract

As assessment methods are changing, the way to determine their quality needs to be changed accordingly. This chapter argues for the use Competence Assessment Programmes (CAPs), combinations of traditional tests and new assessment methods which involve both formative and summative assessments. To assist schools in evaluating their CAPs, a self-evaluation procedure was developed, based on 12 quality criteria for CAPs developed in earlier studies. A self-evaluation was chosen as it is increasingly used as an alternative to external evaluation. The CAP self-evaluation is carried out by a group of functionaries from the same school and comprises individual self-evaluations and a group interview. The CAP is rated on the 12 quality criteria and a piece of evidence is asked for to support these ratings. In this study, three functionaries from eight schools (N = 24) evaluated their CAP using the self-evaluation procedure. Results show that the group interview was very important as different perspectives on the CAP are assembled here into an overall picture of the CAP's quality. Schools seem to use mainly personal experiences to support their ratings and need to be supported in the process of carrying out a self-evaluation.

[5] This chapter is based on:
Baartman, L. K. J., Prins, F. J., Kirschner, P. A., & Van der Vleuten, C. P. M. (2007). Determining the quality of Competence Assessment Programmes: A self-evaluation procedure. *Studies in Educational Evaluation*, *33*, 258-281.

Introduction

Education is undergoing fundamental changes in many European countries. In the Netherlands, new qualification structures for vocational education have been developed which are based on competences and work-related experiences. The rationale behind these innovations is to better link educational programmes to job requirements and to enable vocational education to incorporate new developments taking place in the work field (Tillema, Kessels, & Meijers, 2000). From 2010 on, Dutch vocational institutions are legally bound to adopt a competence-based curriculum, focusing on the competences (knowledge, skills and attitudes) needed in relevant job situations. As an important part of education, assessment is changing as well (Birenbaum, 1996; Dochy & McDowell, 1997). Competence-based curricula require different assessment approaches to adequately determine competence acquisition. As competence can be seen as the capacity to enact specific combinations of knowledge, skills and attitudes in appropriate job contexts (Lizzio & Wilson, 2004), assessment should focus on the integration of these three elements. This implies that in addition to assessing content knowledge, skills and attitudes should be assessed, and this should be done in an integrated way.

In the transition towards assessment of competence, assessment quality has played a key role. Traditional knowledge-focused assessment approaches are currently being criticised by a number of researchers. Recently, Birenbaum et al. (2006) stated that traditional assessment approaches focus on assessment of learning instead of assessment for learning, are limited in scope and ignore individual differences increasingly encountered in education. Although part of this might be true, alternative assessment approaches currently being developed are not without problems either. Though they are supposed to be more valid than traditional assessments (Birenbaum, 1996; Linn, Baker, & Dunbar, 1991), some feel that the evidence against traditional tests is not as strong as has been claimed, and that the claim that newer forms of assessment are more valid and suitable still needs empirical confirmation (e.g., Glaser & Silver, 1994; Hambleton & Murphy, 1992; Messick, 1994). This chapter does not attempt to resolve the dispute between traditional and new approaches to assessment. Instead, we argue that (1) it is unwise to assume that new approaches to assessment are a panacea for solving all assessment problems, and (2) traditional and new assessments can be viewed as playing complementary rather than contradictory roles (Birenbaum, 1996; Maclellan, 2004). Therefore, in earlier publications (Chapter 2; Baartman, Bastiaens, Kirschner, & Van der Vleuten, 2007a) we proposed the use of Competence Assessment Programmes (CAPs), which are defined as combinations of traditional and new forms of assessment in an assessment programme, which can have both formative and summative functions.

Programme quality versus single method quality

There are a number of reasons why it is important to think in terms of programmes of assessment and why the quality of such a programme should be evaluated as a whole. First, since competence involves the integrated application of knowledge, skills and attitudes, it is often argued that one single assessment method is not enough to assess competence and that a mix of methods should be used instead (e.g., Chester, 2003; Van der Vleuten & Schuwirth, 2005). Second, Knight (2000) argues for a programme-wide approach to assessment in which attention is concentrated upon all assessment arrangements in complete educational programmes. The advantage of this approach is that the reliability pressure on low-stake assessments in a programme can be reduced and the resources freed up can be invested in the development of costly and reliable (and more valid!) assessments where they are needed, namely in high-stake situations. Third, a CAP comprises assessments with both formative and summative purposes. The main functions of formative assessment are providing feedback and generating appropriate learning activities, whereas summative assessment mainly serves to enable grading decisions (Black & Wiliam, 1998; Gibbs, 1999). Therefore, Knight (2000) argues for the need to make an explicit distinction between formative and summative assessment, in which reliability is less important for formative assessment and where summative assessments should be made as reliable as possible. Although he does urge not to diminish the validity of summative assessments, we feel that making such a clear distinction between formative and summative assessments runs the risk of evaluating formative assessment on new, learning-related criteria, and summative assessment on traditional, technical criteria. As summative assessments also have a 'formative potential' (Hickey, Zuiker, Taasoobshirazi, Schafer, & Michael, 2006) in steering students' learning processes, we argue that learning-related quality criteria are just as important for summative assessments and that CAP quality should be evaluated integrally.

To evaluate CAPs, this chapter uses 12 quality criteria developed and validated in earlier studies (Chapters 2, 3, and 4; Baartman et al., 2006; 2007a, 2007b). Table 5.1 lists the quality criteria and gives a short summary of each. The rationale behind the quality criteria is that since CAPs consist of both traditional and new forms of assessment, both traditional and new quality criteria are needed to evaluate their quality. Our previous work (Chapter 2; Baartman et al., 2007a) addressed some problems with regard to the use of reliability and validity for CAPs and suggests operationalising reliability and validity in a different way and complementing them with other quality criteria proposed for new forms of assessment, such as the *consequences*, *meaningfulness* and *cognitive complexity* of an assessment (e.g., Kane, 1992, 2004; Linn et al., 1991; Van der Vleuten & Schuwirth, 2005).

Table 5.1: Short description of the twelve quality criteria for CAPs

Criterion	Short description
Acceptability	All stakeholders should approve of the assessment criteria and the way the CAP is carried out. They could have confidence in the CAP's quality
Authenticity	The degree of resemblance of a CAP to the future workplace, in terms of the assessment task, the physical and social context, and the assessment criteria
Cognitive complexity	A CAP should reflect the presence of the cognitive skills neeeded and should enable the judgment of thinking processes
Comparability	CAPs should be set up and carried out in a consistent way. The tasks, criteria and working conditions should be consistent with respect to key features of interest, and scoring should occur in a consistent way
Costs and efficiency	The feasibility of developing and carrying out the CAP for both students and assessors, and the time and resources needed, compared to the benefits
Educational consequences	The degree to which the CAP yields positive effects on learning and instruction, and the degree to which negative effects are minimised
Fairness	Students should get a fair chance to demonstrate their competences, for example by letting them express themselves in different ways and by making sure the assessors do not show biases
Fitness for purpose	Alignment among standards, curriculum, instruction, and assessment. The assessment goals and methods used should be compatible with the educational goals
Fitness for self-assessment	CAPs should stimulate self-regulated learning. They should include specific methods to foster this learning such as practice in self-assessment and giving and receiving feedback
Meaningfulness	CAPs should have a significant value for all stakeholders involved. For learners, assessments should be a learning experience in themselves, and be useful for the learning process. For teachers and employers, the assessments should be meaningful in terms of the requirements of the future job
Reproducibility of decisions	The decisions made on the basis of the results of CAP should not depend on the assessor or the specific assessment situation. Therefore, multiple assessors, assessment tasks, and situations should be combined
Transparency	CAPs should be clear and understandable to all stakeholders. Learners and assessors should know the scoring criteria, and the purpose of the assessments. External controlling agencies should be able to get a clear picture of the way in which a CAP is developed and carried out.

Very little is known about how to determine the quality of an assessment programme instead of the quality of a single assessment method. Stokking, Van der Schaaf, Jaspers, and Erkens (2004) state that the criteria used should depend on whether the assessment is used formatively or summatively. For formative assessments, *comparability* and *reproducibility* can get less priority, whereas efficiency is very important to assure that feedback can be given often and efficiently. For summative assessments, special measures to assure *comparability*, *reproducibility* and *fairness* should be a standard procedure. Transferring these ideas to programme quality implies that not all single assessment methods in a CAP must meet all quality criteria. Although we can be more lenient with regard to *reproducibility* and *comparability* for formative assessment, the summative assessments within a programme should comply with all quality criteria, including learning- and feedback-related criteria like *meaningfulness* and *educational consequences*. A CAP as a whole has to comply with all quality criteria. For example, high scores on *authenticity* cannot offset major deficits in *cognitive complexity*.

This research focuses on the evaluation of assessment programmes. Many European countries are currently developing competence-based educational programmes and concomitant assessment programmes. In the United States, a similar movement towards what is called performance standards-based education can also be observed (Valli & Rennert-Ariev, 2002). These assessment programmes often consist of combinations of traditional and new assessment methods. Our goal is to explore whether the quality of these programmes as a whole can be determined and whether schools can do so using a self-evaluation method developed for this study.

School Self-Evaluations

Assessment quality can be demonstrated in a large number of ways, of which self-evaluation is just one. Jonsson, Baartman, and Lennung (submitted), for example, evaluated an assessment programme by means of analyses of student examination scores and student questionnaires, and in many countries external auditing is a commonly used method to assure assessment quality. A self-evaluation method was chosen here because in many European countries, school self-evaluation is becoming an increasingly important approach to both school improvement and accountability (McNamara & O'Hara, 2005). School self-evaluation or internal evaluation is carried out by a school itself, for example by a group of teachers, the department or school manager, a specific staff member, or a combination thereof. In contrast, external evaluation is carried out by someone outside the school, usually inspectors or governmental organisations, and mainly serves accountability purposes (Nevo, 1994, 2001). In these discussions about internal and external evaluation, school improvement and self-evaluation refer to the educational process as a whole, and not specifically to assessment.

In many countries there is a movement towards pulling back direct government involvement in day-to-day activities (i.e., fewer rules, deregulation, decentralisation towards municipalities, and a wider scope for schools to pursue

their own policy) and replacing this with more school autonomy and the requirement that the schools make their own policy and prove that they have met the governmental requirements. In the Netherlands, for example, there has been a movement over the last decade to increase school autonomy, which is counterbalanced by more centralisation in the areas of curriculum and outcomes assessment (Scheerens, Van Amelsvoort, & Donoghue, 1999). For assessment specifically, self-evaluation has become a topic of debate in the Netherlands since vocational institutions have to demonstrate the quality of their assessments to an external quality board (EQC: Examination Quality Centre) in order to retain their accreditation. In this model, schools carry out self-evaluations, which serve as a starting point for the external evaluations carried out by the EQC. This line of development is described by Kyriakides and Campbell (2004) as a progressive line of maturation of the school system from a controlling external inspection to more co-operative models in which internal and external evaluation co-exist.

Studies on the use of self-evaluation have shown positive results. Teachers appear to be willing to be self-critical and experience self-evaluation as less threatening than external evaluation (McNamara & O'Hara, 2005). They reported that what they learned from the self-evaluation had a significant impact on their teaching and their professional perceptions and behaviour (Nevo, 1994). When carrying out self-evaluations, schools are more self-confident and less defensive when confronted with negative findings from external evaluation (Nevo, 2001). Evaluation is thought to be most effective when people internalize quality standards and apply them to themselves, as they do in self-evaluation (McNamara & O'Hara). Difficulties reported with regard to self-evaluation are the need for significant resources and skilled personnel (Nevo, 2001), the often encountered judgment of low validity and reliability (Scriven, 1991), and the lack of sufficient and appropriate data and evidence to support the school's claims about their strengths and weaknesses (McNamara & O'Hara).

In sum, research on self-evaluation shows the merits of self-evaluation, but also some possible pitfalls, one of which is the fact that schools often do not support their claims by using appropriate evidence. Previous studies, though, have not looked into the exact nature of the support presented in self-evaluations. This study does look at this support from the perspective of argumentation theory and takes a qualitative approach to gain a deeper understanding of the processes taking place during self-evaluation. Research on argumentation shows that the ability to provide support for one's claims cannot be taken for granted (Kuhn, 1994). If self-evaluation is to be a valuable approach to both school improvement and accountability, a precondition is that schools are capable of performing self-evaluations. In this study, a self-evaluation procedure was developed to assist schools in evaluating their newly developed CAPs, based on the 12 quality criteria for CAPs developed and validated in earlier studies (Chapters 2, 3, and 4; Baartman et al., 2006, 2007a, 2007b). Eight vocational schools participated, and in each of these schools, three functionaries collaboratively evaluated their CAP using the self-evaluation procedure. We explored whether they are capable of evaluating their own CAP and

whether they can support their claims by means of examples or evidence (i.e., whether they could substantiate their claims). The CAP quality self-evaluation procedure is described in the next section. The method section that then follows describes how we went about evaluating the self-evaluation procedure in this study.

The CAP quality self-evaluation procedure

The goal of the self-evaluation method developed here is to stimulate schools to reflect on the quality of their CAP and to provide ways to improve this CAP. As such, it has no summative goal and has no consequences as has an audit by the EQC. As it is meant to evaluate assessment programmes - not single assessments - the evaluators need to have an adequate overview of all assessments used within the programme. This could be a programme for a specific year (e.g., an introductory year), for a specific subject area (e.g., biology) or even for an entire educational programme (e.g., a nursing programme). Few people within a school probably have this overview and therefore the self-evaluation method requires a group of personnel from the same school (e.g., year, domain, programme) to collaboratively evaluate their own CAP. The self-evaluation method consists of two phases. First, all users individually evaluate their CAP using a web-based self-evaluation tool. In the second phase, all individual evaluations are assembled and discussed in a group interview.

Phase 1: Individual CAP self-evaluations

The individual self-evaluations of a school's CAP are carried out with a web-based evaluation tool, based on the 12 quality criteria. Before evaluating their CAP, the evaluators are asked to describe it by indicating the year(s) and level of education, and the assessment methods included. Examples of methods are given, including multiple choice test, written test with open questions, presentation, assessment of products made, assessment interview, criterion-based interview, observation in a simulated situation, observation in the workplace, portfolio and proof of competence. Additional forms of assessment could be added by the user.

Subsequently, they evaluate their CAP on the 12 quality criteria for CAPs developed earlier. For the self-evaluation tool, these quality criteria are operationalised as indicators: more concrete aspects of a quality criterion in practice, though not too detailed so that they turn the self-evaluation into just ticking off a checklist. Per quality criterion, four to six indicators were formulated, based on a literature study and a previously carried out pilot study (e.g., Baume, Yorke, & Coffey, 2004; Benett, 1993; Dierick & Dochy, 2001; Dochy, Gijbels, & Van de Watering, 2004; Gulikers, Bastiaens, & Kirschner, 2004; Linn et al., 1991; McLellan, 2004; Miller & Linn, 2000; Moss, 1994; Schuwirth & Van der Vleuten, 2004). Along with the pre-determined indicators, two open fields are included for each criterion, so that users can include more and other indicators relevant to their situation. Table

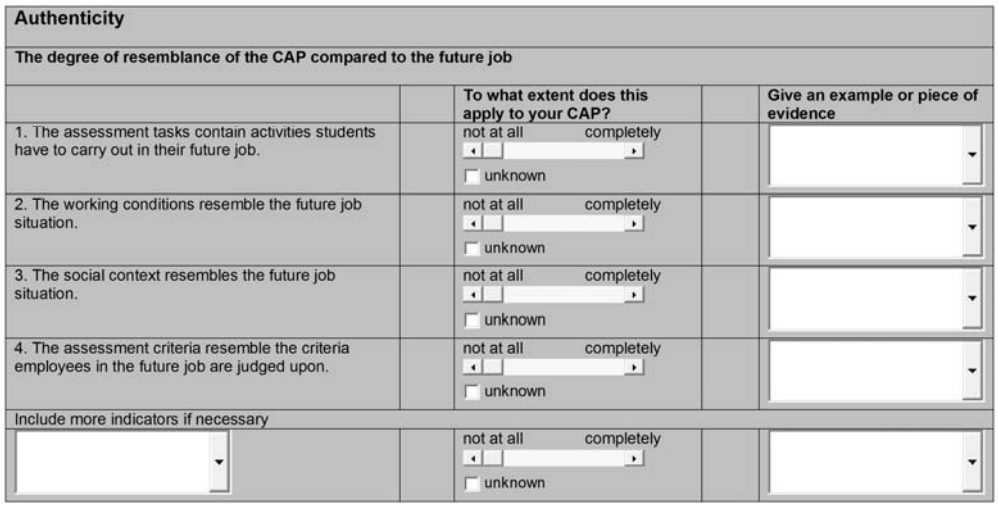

Figure 5.1. Schematic representation of the web-based self-evaluation tool

5.2, on page 101 in the methods section, gives an overview of all quality criteria and an abbreviated version of their indicators.

The CAP is evaluated both quantitatively and qualitatively. For the quantitative evaluation, the CAP self-evaluation tool asks the evaluators to rate the CAP on each indicator via an analog slide-bar that can be moved from not at all to completely (see Figure 5.1). A don't know option was available as well. Behind this slide bar is a rating scale ranging from 0 to 100, which is invisible so as not to give evaluators the idea of giving a score or mark to their CAP. For the qualitative evaluation, the tool asks for support of the ratings given in the form of an example or evidence showing that the CAP indeed complies with the indicator. The self-evaluation tool is complemented by an instruction page and a vocabulary list in which all different assessment methods are defined and explained. The instructions and vocabulary list can be accessed at any time. Figure 5.1 presents a schematic representation of a page of the web-based self-evaluation tool.

Phase 2: Group interview

After the individual CAP self-evaluations, all individual ratings and support thereof are assembled and collected in an overview of the school's CAP quality. For each quality criterion, the overview presents the ratings and support of the indicators given by all evaluators. The overview is used as input for the group interview, which is meant to stimulate discussion and reflection on CAP quality and to result in an overall picture of the quality of the CAP. The group interview lasts about two hours and has a semi-structured character. First, the evaluators are asked to globally describe their CAP. They are given the list of assessment methods they ticked off in the self-evaluation tool and are asked to describe them more elaborately and to indicate the percentage of total assessment time devoted to each. Second, the overview with all evaluators' ratings and support is discussed and they are explicitly

encouraged to comment on their own and each others' ratings and support. If they change their minds about a rating or support thereof during the group interview, they are allowed to adjust their initial rating and/or support given in the individual self-evaluation (comparable to a Delphi-study approach). This is noted down by the interviewer, who asks for further information or explanation if the: (1) argumentation is unclear to the interviewer, and (2) evaluators have clearly different opinions. To get an indication of 'a clearly different opinion' the range of ratings was divided into three categories: 0-35 (low), 36-65 (medium), and 66-100 (high). A clearly different opinion was operationalised as falling in different categories and differing at least 20 points. To conclude the group interview, the evaluators are asked to collaboratively summarise the strong and weak aspects of their CAP, based on the individual self-evaluations and the group interview.

As stated, the purpose of this study is to explore whether schools are capable of evaluating their assessment programme using the CAP-quality self-evaluation procedure. A pilot study was carried out, in which two school managers, three teachers, two examination board members, and two EQC auditors carried out the self-evaluation and were explicitly asked to comment on the clarity and understandability of the quality criteria and indicators. Most quality criteria were found to be clear and understandable. Unclear indicators or indicators found to be too abstractly formulated were reformulated for this study. The research questions of this study focus on the process of carrying out the self-evaluation, and not on the product of it, that is, if the CAPs evaluated are of sufficient quality. One specific aspect we studied is the evidence that the participants gave to support their ratings, which was explored in a qualitative way from the perspective of argumentation literature. Research questions are: (1) How do the two parts of the self-evaluation procedure, that is the individual phase and the group interview, contribute to both the process and the outcomes of the self-evaluation? and (2) What are the nature and the quality of the support given to the ratings?

Method

Context of the study

This study was carried out in Laboratory Technology Education in vocational institutions in the Netherlands. Within the Dutch educational system, after leaving primary schools, all pupils are required to enter secondary education where they can choose between general secondary education which leads to university studies, or polytechnics, and pre-vocational education (age 12-16). Pre-vocational education serves as a preparation for upper secondary vocational education (age 16-20). Laboratory Technology is a vocational programme preparing students for a job as laboratory assistant or laboratory technician. The schools participating in this study were organised in a national consortium of vocational schools that started to

implement problem-based education in 2000/2001 and is now working towards competence-based education.

Participants

Laboratory technology departments of eight vocational schools participated in this study. At each school, the department manager, a member of the examination board and another teacher participated. The pilot study carried out earlier in a different school revealed that these three functionaries generally are acquainted with the assessments used in the department. Together they have a full overview of all assessments used, both from the point of view of policies and regulations and from practical experience. The ratings and support of two participants were left out of the analyses. These participants, one teacher and one examining board member from two different schools, did not have enough insight into their school's CAP to carry out the individual self-evaluations. They acknowledged this themselves at the start of the group interview, but did participate in the interview to gain more insight into their CAP.

Procedure

All schools were contacted through the national consortium of vocational schools, within which the laboratory schools are organised as a content-specific working group. One week before the group interview, all participants received an email asking them to independently fill out the web-based self-evaluation tool. The three participants from each school were asked to first collaboratively determine the CAP they would evaluate, for example all assessments used in the first year of the educational programme. This ensured all participants from one school had the same CAP in mind. The participants then individually used the CAP self-evaluation tool to evaluate the chosen CAP. The group interviews were carried out by the first author approximately a week later and lasted about 2 hours. At the start of the group interview, the participants were presented with the overview of all individual CAP self-evaluations. All interviews were audio taped with permission of the participants.

Data analysis

To answer the first research question on the contribution of the individual phase and the group interview to the processes and outcomes of the self-evaluation, both quantitative and qualitative analyses were carried out. First, the percentage of ratings completed with a piece of support was calculated before and after the interview. The ratings given were divided into the three categories used in the group interview: low (0-35), medium (36-65), and high (66-100), together with a don't know category. The percentages of low, medium, and high ratings per indicator given before and after the interview, and the changes made during the interview (e.g., a change from a low rating to a high rating) were calculated. Second, all group interviews were transcribed literally and analysed qualitatively. The first author

analysed the group interviews by noting recurrent themes, for example if the participants had thought of a specific part of their CAP instead of the entire CAP when giving their ratings and support. The first themes were identified when analysing the first interviews. The other interviews were used to check whether they could be found again and new and different themes were added to the first ones. This process continued until no new themes could be identified, which was the case after analysing all interviews three times. Then, the list of themes found by the first author was given to a researcher not involved in the current project who independently analysed the group interviews. She identified the themes listed by the first author by marking the parts of the transcribed interviews belonging to each theme and added new themes to the list made by the first author. The first author and the independent researcher discussed the themes until agreement over the list was reached.

To answer the second research question on the nature and quality of the support, Miles and Huberman's (2003) phases for qualitative data analysis were followed, in which qualitative data are first meaningfully reduced or reconfigured (data reduction), then organised into different data displays such as diagrams and matrices (data display), from which conclusions can be drawn and verified in the last phase (conclusion and verification). In the first phase of analysis, a summarising display was constructed for each school with the ratings and support given in the individual CAP self-evaluations, complemented with the ones found in the typed out group interviews. The support was summarised over the three participants per school, resulting in an overview of the support given for each school. If the participants agreed on the support given, this was summarised in the overview. If they did not agree, two or three different pieces of support were included in the analysis. For the second phase of analysis, the eight overviews for the separate schools were assembled in a meta-matrix. The support was now summarised over schools, resulting in a so-called ordered matrix including all different pieces of support together with their ratings, which were again categorised into low, medium and high. From this ordered meta-matrix, conclusions were drawn in the last phase of analysis. To assure the qualitative analyses did not depend on the authors' personal and subjective interpretations, a check (verification) was carried out by a researcher not involved in the current project who independently re-constructed the data displays. Only very small differences between the two researchers were found, which were changed in accordance with both researchers' opinions. To assure further verification, the first and second author together carried out the final conclusion phase, for which a flow chart for coding the quality of arguments developed by Clark and Sampson (2005) was adjusted for this research. Clark and Sampson's flow chart is based on Toulmin's (1958) well-known scheme of the layout of arguments. In argumentation literature, some researchers analyse the quality of argumentation by investigating if every element of Toulmin's scheme is present (e.g., Simon, Erduran, & Osborne, 2006), but Clark and Sampson argue that these analyses should also include judgments of the quality of the arguments, and not just their absence or presence. The flow chart classifies the quality of argument as either: no

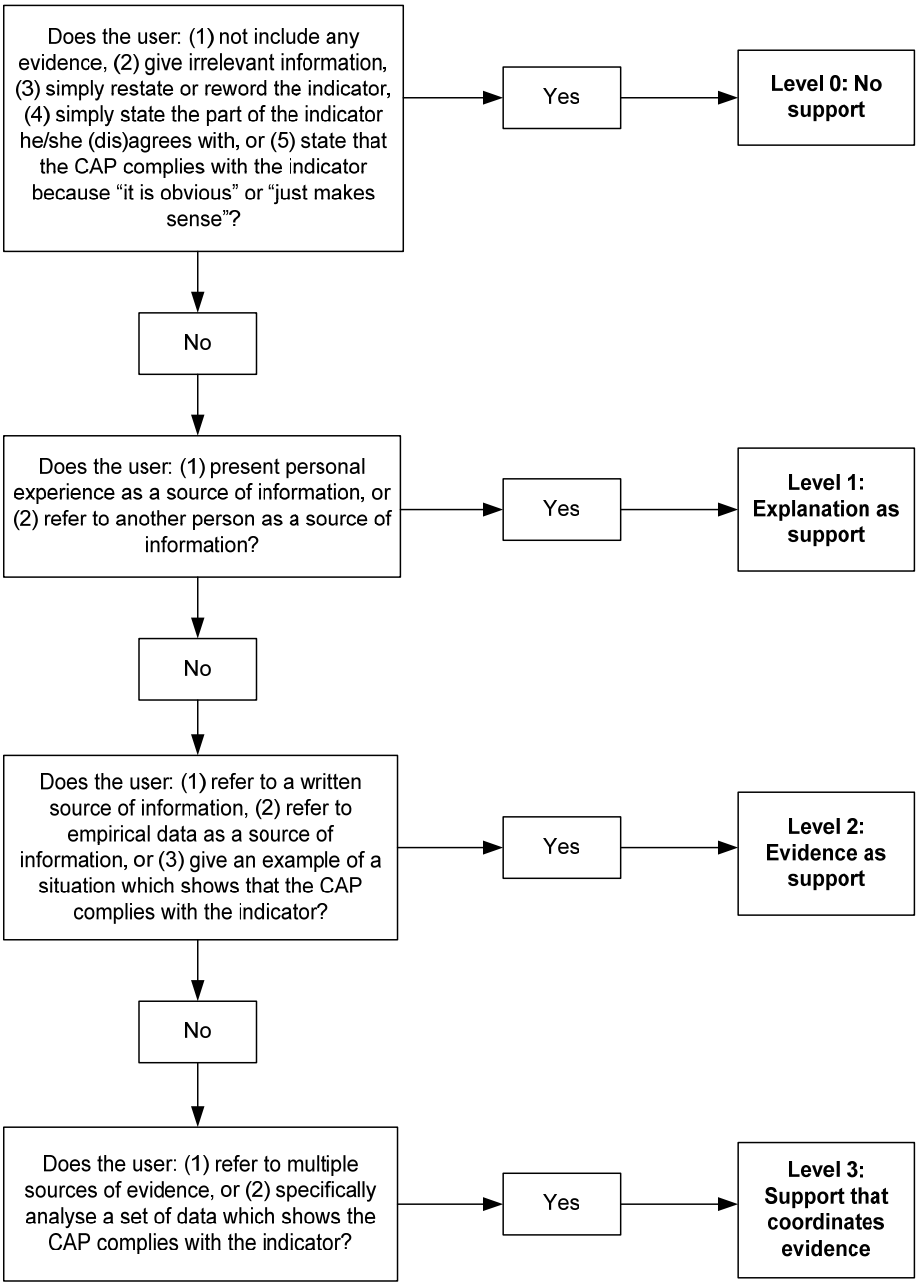

Figure 5.2. Scheme for analysing the nature and quality of support

support (level 0), using explanation as support (level 1), using evidence as support (level 2) and coordinating multiple pieces of evidence or multiple connections between ideas in the evidence (level 3). The flow chart was adjusted by referring to the quality of CAPs instead of arguments in a group discussion. Figure 5.2 presents the adjusted flow chart used in this research. The first and the second author independently coded all support using the flow chart and kappa values were calculated to check for interrater reliability. After coding the argumentations of twoquality criteria, the initial interrater reliability was found to be mediocre to good (.51 and .70). The different codings were discussed and the largest differences between the two researchers appeared to involve the distinction between level 0 and level 1. From some pieces of support, it did not become completely clear whether the participant was really adding any new information, or was merely repeating the indicator. It was decided to score the indicator as level 0 when it was not completely clear what the participant was exactly referring to and whether this could be considered as additional information to the indicator, although additional information might have been present implicitly. After resolving these differences, all support was scored and interrater reliabilities were found to be satisfactory (Cohen's kappa ranging from .70 to .87). The codes from the second author were used for further analyses, as the first author conducted all interviews and might thus be biased towards certain schools.

Results

Before presenting the results with regard to the two research questions posed, the Cronbach's Alpha values of the criterion scales of the 12 quality criteria are discussed here. Although we did offer the possibility to include other indicators than the ones proposed and we do not pretend to give a full overview of all possible indicators, the indicators were designed as a scale of each quality criterion. Table 5.2 presents the Cronbach's Alpha values found for the criterion scales (in bold) and the item-total correlation for each indicator. Taking .60 as an acceptable Alpha value, six criteria initially could not be considered as a scale. In addition, a number of indicators had low item-total correlations. These correlations should be higher than .35, but lower values are accepted if items cannot be missed theoretically.

A reason that may explain some low Alpha values is that Cronbach's Alpha increases when the sample size increases and when the number of items within a scale increases. In this study, we had a relatively small number of participants (N = 22), and some indicators had many missing values. In this case, the missing values are the percentages in the don't know category in Table 5.2. As any unclear indicators were explained during the group interview, a high percentage don't know seems to indicate that the participants indeed did not know whether their CAP complied with the indicator or not, and not whether they understood the indicator or not. For the Alpha values this resulted in a lower sample size, which was sometimes reduced by almost half. For this reason, the Alpha values were re-calculated using mean substitution for the criteria with an insufficient Alpha value.

This resulted in an acceptable Alpha value for the quality criterion *acceptability*. The other quality criteria still had insufficient Alpha values, which necessitated the deletion of indicators. The last column of Table 5.2 shows the re-calculated Alpha values with deletion of the indicators with the lowest item-total correlations. The re-calculated Alphas are sufficient, although the *reproducibility* and *transparency* scales need to be interpreted with some caution. Although statistically the deletion of indicators was necessary, at this first stage of development and use of the indicators and scales we are reluctant towards permanently deleting indicators. At this stage of development of CAPs it is very well possible that schools do pay attention to one indicator, but not to another. For example, for *acceptability* some schools may have asked students' opinions, but not those of the employers, and other schools may have done so the other way around. This difference results in a low Alpha value, which does not mean that the theoretical concepts within the scale do not fit together. In sum, at this point we had to delete some indicators from the scales and some indicators indeed may not fit in a scale theoretically, but further research using larger samples is needed before final conclusions can be drawn here. Moreover, some indicators may get less don't know answers in the future, when schools are more used to the newer ideas of the quality of assessment presented in the indicators. At this point, we will present the results of this study on the scale level as much as possible, but we will refer to the indicators when necessary.

The individual self-evaluations and group interviews

With regard to the contribution of the individual self-evaluations and the group interview to the school self-evaluation process, two categories of results are presented. The first category involves the ratings and support given before (the individual self-evaluations) and after the group interview, and the changes made during the group interview. The second category involves the categories of recurrent themes observed during the group interviews.

Ratings and support before and after the group interview

The first two columns of Table 5.2 present the percentages of ratings completed with a piece of support before and after the interview. As can be seen, before the group interview 63% of the ratings were supported with a piece of evidence. After the interview this percentage had increased to 76%, meaning that support was added or complemented during the interview. As the data are non-parametric, this difference was tested by means of a Wilcoxon's signed ranks test and was found to be significant ($Z = -6.182$, $p < .001$). In total, 58 ratings were changed during the group interview. Most changes were from a low rating to a high rating (15), followed by changes from a low to a medium rating (9) and a low rating to a don't know (9). Apparently, the group interview and the discussions among the participants caused them to give higher ratings than they had initially given individually before the interview. The changes from a low rating to a don't know were mostly caused by the fact that the participants realized they had given a rating without being able to.

Table 5.2. Ratings and support given before and after the group interview[1]

		Before interview	After interview						
Criteria and Indicators		**% Subst.**	**% Subst.**	**% low ratings (0-35)**	**% med ratings (36-65)**	**% high ratings (65-100)**	**% don't know**	**α & Item-Total**	**re-cal. α & Item-Total**
Acceptability		66	78	11	21	43	25	.51	.65
1	Students approve of criteria	59	82	9	14	59	18	.28	.44
2	Students approve of procedure	55	73	9	32	41	18	.87	.43
3	Teachers approve of CAP	73	82	14	27	55	5	-.56	.27
4	Employers approve of CAP	55	86	9	14	18	59	.38	.25
5	Confidence in quality CAP	59	68	14	18	41	27	.61	.69
Authenticity		76	85	15	23	61	1	.70	
1	Assessment tasks resemble job	82	82	0	9	91	0	.23	
2	Working conditions resemble job	82	91	18	41	41	0	.56	
3	Social context resembles job	68	82	27	23	50	0	.64	
4	Assessment criteria resemble job	73	86	14	18	64	5	.55	
Cognitive complexity		65	70	30	24	36	10	.74	
1	Tasks trigger thinking steps	68	73	14	27	41	18	.64	
2	Explain choices	68	73	41	18	32	9	.47	
3	Criteria address thinking steps	55	59	41	23	23	14	.72	
4	Tasks require thinking level	68	77	23	27	50	0	.36	
Comparability		65	72	5	8	85	2	.72	
1	Assessment tasks comparable	77	86	9	9	82	0	.18	
2	Working conditions comparable	59	68	0	18	77	5	.65	
3	Assessment criteria comparable	64	64	0	5	91	5	.71	
4	Assessment procedure comparable	59	68	9	0	91	0	.59	

		Before interview	After interview						
Criteria and Indicators		**% Subst.**	**% Subst.**	**% low ratings (0-35)**	**% med ratings (36-65)**	**% high ratings (65-100)**	**% don't know**	**α & Item-Total**	**re-cal. α & Item-Total**
Costs and efficiency		56	69	26	18	34	22	.41	.69
1	Time and money estimated	55	73	45	14	23	18	.44	.67
2	Deliberately choosing mix	55	68	32	14	36	18	.42	.69
3	Yearly evaluation of efficiency	59	73	18	23	45	14	.35	.23
4	Positive effects outweigh investments	55	64	9	23	32	36	-.25	
Educational Consequences		65	64	18	22	42	18	.46	.71
1	Desired learning processes stimulated	64	73	32	23	41	5	.63	.52
2	Positive influence on students	59	73	18	23	27	32	.25	.43
3	Positive influence on teachers	55	68	18	27	27	27	-.05	.57
4	Improved if negative effects	77	86	5	14	82	0	.36	
5	Curriculum adapted if CAP warrants	73	64	18	23	32	27	.17	.49
Fairness		60	75	7	15	63	15	-.44	.77
1	Procedures to rectify mistakes	59	73	0	18	59	23	.05	.57
2	Weights based on importance	68	82	32	14	41	14	-.42	
3	Assessors not prejudiced	59	77	5	23	64	9	.26	.81
4	Various types of assessment tasks	45	64	0	14	77	9	-.38	
5	Student think CAP is fair	68	77	0	5	73	23	.25	.51
Fitness for Purpose		68	85	17	16	65	1	.70	.79
1	Coverage of competence profile	77	95	0	23	77	0	.78	.78
2	Integrated assessment of K/S/A	77	95	41	27	32	0	.58	.64
3	Mix of different assessment forms	59	77	0	9	91	0	-.32	

		Before interview	After interview						
Criteria and Indicators		**% Subst.**	**% Subst.**	**% low ratings (0-35)**	**% med ratings (36-65)**	**% high ratings (65-100)**	**% don't know**	**α & Item-Total**	**re-cal. α & Item-Total**
4	Both summative and formative forms	64	77	32	0	68	0	.48	.50
5	Forms match with educational goals	64	82	14	23	59	5	.74	.72
Fitness for Self-Assessment		61	69	31	23	40	7	.86	
1	Self- and peer-assessment	73	95	18	27	55	0	.49	
2	Giving and receiving feedback	59	68	23	32	41	5	.68	
3	Reflection on personal development	55	55	32	18	41	9	.92	
4	Formulation of personal learning goals	59	59	50	14	23	14	.75	
Meaningfulness		51	68	24	19	33	25	.93	
1	Feedback formative useful	55	86	18	14	41	27	.83	
2	Feedback summative useful	64	95	23	23	27	27	.87	
3	Assessment is opportunity to learn	41	55	41	23	18	18	.81	
4	Students think criteria meaningful	41	50	23	18	27	32	.68	
5	Teachers/employers think criteria meaningful	55	55	14	18	50	18	.88	
Reproducibility of decisions		63	85	27	27	40	6	.38	.59
1	Several times	68	91	18	32	41	9	.04	
2	Several assessors	68	86	9	23	59	9	.36	.41
3	Assessors with different background	64	86	55	14	27	5	.13	.24
4	Equal discussion between assessors	64	86	23	18	50	9	.23	.41
5	Trained and competent assessors	55	86	32	45	23	0	.07	.28
6	Several work situations	59	73	23	32	41	5	.42	.46

		Before interview	After interview						
Criteria and Indicators		**% Subst.**	**% Subst.**	**% low ratings (0-35)**	**% med ratings (36-65)**	**% high ratings (65-100)**	**% don't know**	**α & Item-Total**	**re-cal. α & Item-Total**
Transparency		65	76	9	25	57	9	.43	.58
1	Student know formative or summative	77	86	0	23	73	5	.38	.47
2	Students know criteria	59	73	9	41	36	14	-.06	.36
3	Students know procedures	59	64	5	32	55	9	.49	.51
4	Teachers know and understand	73	86	0	27	68	5	.27	.26
5	Employers know and understand	64	82	23	9	50	18	.36	.21
6	External party can audit	59	64	18	18	59	5	.01	
Total		63	76	18	20	50	12		

[1.] *The indicators in this table are summarised for practical space reasons. A full description of all indicators can be obtained from the first author*

support it: 'Actually, I don't have any experience with assessing choices at this moment … I should have put 'don't know'' [School 1]. It needs to be remarked here that in total very few changes in ratings were made during the interview. In total 1254 ratings were given, of which 58 were changed (4.6%). Apparently, the interviews had a greater effect on the support than on the ratings given

In addition, Table 5.2 presents the percentages of low, medium and high ratings given after the interview. The ratings after the interview were taken here to include any 'corrections' made and because few changes were made at all during the interview. In total, many more high (M = 50%) than low (M = 18%), medium (M = 20%) and don't know ratings (M = 12%) were given. Friedman's non-parametric test showed that the differences between the percentages of low, medium and high ratings was significant (χ^2 = 72.727, p < .001). Wilcoxon's signed ranks tests showed that the differences between the percentage of high and low ratings and the difference between the percentage of high and medium ratings were significant (Z = -5.501, p = .000 and Z = -6.142, p < .001 respectively). The difference between the percentage of low and medium ratings was found to be non-significant (Z = -1.453, p = .146).

Apparently, the participants gave their CAP relatively high ratings. The highest percentage of high ratings was found for *comparability* (85%). This is a quality criterion that traditionally has been paid much attention to, and this does not seem to have decreased during the transition towards competence-based education. The lowest percentage of high ratings were found for *meaningfulness* (33%), *costs and efficiency* (34%), and *cognitive complexity* (36%). These quality criteria are newer and schools may be less familiar with these concepts.

Recurrent themes in the group interview

A list of seven recurrent themes was extracted from the group interviews, which can be categorised into three groups of related themes which are further elaborated in the next sections.

Rating and supporting the indicators:

1. The participants give ratings and support for a broader CAP than agreed upon;
2. The participants give ratings and support for a specific smaller part of the CAP;
3. The participants describe how they would like their CAP to be, instead of rating and supporting the actual situation;
4. The participants say their school is in a transition period towards competence-based education, and therefore some indicators cannot be answered yet or will change in the near future;

The added value of the group interview:

5. The participants perceive their CAP from a different perspective due to their different functionaries within the school, and can therefore complement each other in the group interview;

6. Due to the self-evaluation process and the discussion in the group interview, the participants come up with spontaneous ideas for improving their CAP;

The issue of formative and summative assessment and the audits by the EQC:

7. The participants discuss how to define the formative and summative parts of their CAP and how to present this to the EQC.

Recurrent themes: Rating and supporting the indicators

In the first part of the interview, the participants were asked to shortly describe the different forms of assessment included in their CAP. Here, it became clear that, although they generally agreed on the assessment forms in their CAP, some differences could be observed in how the three participants exactly defined their CAP. As a result, the first part of the group interview tended to serve as a way of collaboratively defining the CAP. When discussing the individual self-evaluations, the participants sometimes appeared to have given a rating for a broader CAP than agreed upon. This was, for example, the case in a school where the participants decided to evaluate their third year's CAP: 'I gave a higher rating, because I only looked at the third year. If you look at the fourth year, for example the proof of competence and the interview … but I didn't include that in my judgment whereas you did' [School 6]. On other occasions, the participants had only thought of a specific part of the CAP when giving a rating: 'Then you're only talking about the summative assessments, I think … I took in mind all assessments' [School 2]. Finally, the participants sometimes appeared to have given a rating based on how they would like their CAP to be, instead of basing their judgment on the actual situation. For example, when discussing *fairness*, this manager said: 'I assume the teachers show professional behaviour … maybe I think they should score 90 here … and people who score lower, they are just not functioning well in their job as a teacher and assessor' [School 1]. These recurrent themes show the participants commented on each others' ratings and support during the group interview and explained their own way of judging the CAP, which contributed to the function of the group interview as a way of 'correcting the mistakes' made during the individual self-evaluations and adding new ratings and support. The last recurrent theme within this category includes the fact that many schools are currently working towards competence-based education and find themselves in a transition period. This also indicates that the ratings and support in this case are likely to change in the near future, when schools have gained more experience with competence-based education and corresponding CAPs.

Recurrent themes: Added value of the group interview

The individual self-evaluations and the group interviews show the department manager, the examination board member and the other teacher perceived their CAP from a different point of view. In the group interview, they tended to complement each other, together creating a more complete picture of the quality of their CAP. Sometimes, the department manager tended to be more negative than the other two

participants because he or she has to deal with complaints from students, teachers and parents, whereas teachers often have both positive and negative experiences in the classroom: 'People who don't agree with the assessment come to me (…) I get the less enthusiastic people. Those who think everything is fine, I don't see' [School 1]. Due to the participants' different functionaries, the group interview often provided the group members with new insights into their CAP, as for example happens in this interview, where the teacher has just told the manager how exactly they go about assessing the students in the laboratory classroom, to which the manager reacts: 'But wow … now I see, that's what I experience right now … you have got a wealth of information about this, also for the audit by the EQC' [School 5]. Finally, the group interview caused the participants to spontaneously come up with improvements for their CAP. For example, when discussing employers' opinions about their CAP, one manager remarked: 'That is difficult to say, but I think it is a good thing the self-evaluation tool asks these questions. It is a signal to us … we have to find out what they think about it' [School 7]. Some other examples are: 'We could specify per assessment project who the assessors are and what influence they have' [School 1], or 'That could be a next step. We could specify and lay down how we want the assessors to carry out the assessment interviews' [School 2].

Recurrent themes: Formative versus summative assessment and the EQC

This final recurrent theme constitutes a content-related issue that came up regularly during both the individual self-evaluations and the group interview. The results show that most schools do not make a clear distinction between formative and summative assessments: 'Well, we don't really make a distinction between formative and summative … what is qualifying and what is part of the learning process' [School 1]. Or: 'At this moment we are still discussing that issue, which assessments to call formative and which summative' [School 6]. This is surprising, as the EQC carries out its audits solely based on the summative assessments and schools have to provide the EQC with an overview of all summative assessments for the audit procedure. Schools experience it as a burden that they have to make a distinction between formative and summative just for these audits: 'We didn't formalize that. We will have to if the EQC comes to visit us, otherwise we have a problem' [School 2]. Or in the words of School 1: 'If the EQC comes, we would call this formative, because otherwise you have to send it all in for the audit, and account for it all. The EQC forces us to condense the summative part.'

Interpreting the results presented so far and looking at how schools define their CAP, give themselves ratings and support these ratings, the preliminary conclusion can be drawn that the group interview was very important in the self-evaluation process. It served to define the CAP as a group and to correct any 'misinterpretations' of the indicators that occurred during the individual self-evaluations. Secondly, it confronted the participants with each others' perspectives, which contributed to obtaining an overall picture of the CAP. Therefore, further analyses on the ratings and their support were carried out on the ratings given after

the interview (thus including any corrections made) and on the support given after the interview (thus including corrections and complementation).

Nature and quality of support

Figure 5.3 presents the percentage of support coded at each level of argumentation distinguished by Clark and Sampson (2005). The main part (in total 56%) of the support was coded as level 1 (explanation as support). Argumentation level 0 (no argumentation) was assigned to 22% of the support and 23% of the support was coded as level 2 (evidence as support). Level 3 (coordinating multiple pieces of evidence) was not found in our data. Support at level 0 was mainly characterized by the fact that it was irrelevant to the indicator at stake or that it was merely a repetition of the indicator. One example is a participant responding to the *comparability* indicator 'assessment procedure comparable' by saying that 'We try to assure all assessment procedures are comparable' [School 3]. Some support at level 0 was characterized by the fact that the participant gave his or her opinion on the matter instead of providing evidence. For example, reacting to the indicator 'assessors with different backgrounds', one participant reacted 'I think the teacher should do the assessments. Students should not have any influence on this' [School 7]. Argumentation level 1 was mainly characterized by participants presenting their

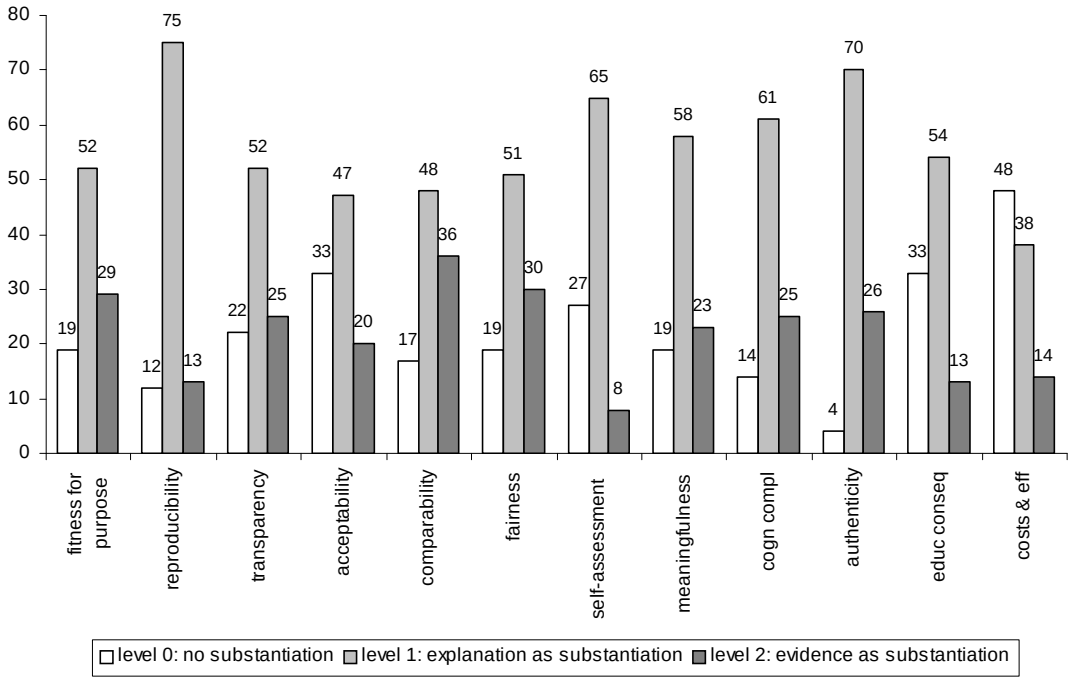

Figure 5.3. Percentages of support within the three levels of argumentation used

own personal experiences, like this participant does for the indicator 'giving and receiving feedback': 'I experience the self-assessment generates valuable feedback on the student's strengths and weaknesses' [School 3]. Support at level 2 involved actual pieces of evidence, for example for the indicator 'improved if negative effects' one manager remarked: 'We recently conducted an evaluation of the assessment and did a brainstorm session with the teachers. We formulated the weaknesses and little groups of teachers are now trying to find solutions to this, for example about how to give better and more immediate feedback' [School 7].

Conclusions and discussion

The purpose of this study was to explore whether schools are capable of evaluating their own Competence Assessment Programme or CAP. A CAP self-evaluation procedure was developed to assist schools in this process. The self-evaluation procedure had a formative function, namely to stimulate reflection on CAP quality and to provide handles for improvement. First, we explored how the two parts of this self-evaluation procedure, the individual self-evaluations and the group interview, contribute to the evaluation of the school's CAP. The results show that the group interview seems to be of great importance. As compared to the individual self-evaluations, support of the ratings was added during the interview. The group interview had less effect on the ratings given, which might be due to the fact that the participants were not explicitly instructed to change their ratings during the group interview, or to reach consensus. In future research and practical use of the self-evaluation procedure for formative purposes, it might be useful to stimulate participants to reach consensus on their CAP's strong and weak aspects and especially on the required improvements, in order to stimulate future use of the results of the self-evaluation for school improvement. The interview also served as a way of collaboratively defining the school's CAP and to correct any 'mistakes' made during the individual self-evaluations. A combination of personnel (in this case the department manager, an examination board member, and another teacher) seems to be useful and necessary if self-evaluation is used for formative purposes. From their different functions within the school the participants add to an overall picture of the school's CAP. Some words of warning are also necessary. First, the fact that two participants had to be left out of the analyses shows that having a good overview of the school's CAP is a prerequisite for being able to evaluate it. Second, the interviews showed that the participants sometimes had difficulties keeping their entire assessment programme in mind during the self-evaluation. Especially when evaluating an entire programme of assessment instead of single assessment methods, schools may need more guidance and instruction. In future research and practical use, it might therefore be useful to include a third phase in the self-evaluation procedure, namely an initial first meeting at the start of the procedure to commonly define the CAP being evaluated.

With regard to second research question about the nature and quality of support, the results showed that the major part of the support given to the ratings

can be categorised as 'explanation as support'. When asked to support their ratings, the participants tended to present their personal experiences, which they used more as a way of explaining why they had given a certain rating, rather than justifying it. This may be due to the fact that the participants were not explicitly encouraged to justify their ratings during the group interview. The self-evaluation tool did ask to support the ratings by a piece of evidence, but the interviewer did not judge or comment on the quality of support given during the interview. Besides, it is important to note that in this study the self-evaluation procedure was formative in character. It had no consequences for the participating schools, as an audit by the EQC has. This may have caused the participants to be more self-critical and to be less focused on justifying their claims, like they have to do for the EQC. Finally, argumentation literature shows that using real evidence to support one's claims is a difficult task that does not come naturally (Kuhn, 1994). Like discussants in a group discussion, schools may need special training to support their claims, and it may be necessary to point out to schools the importance of gathering data on, for example, students' and employers' opinions. At the time of study, almost none of the participating schools possessed any real data on assessment quality specifically, though they usually did evaluate student satisfaction of the entire educational programme.

This study had an exploratory character and focused on the process of carrying out the CAP self-evaluation, and not on the final product of this self-evaluation, that is the actual quality of the CAP being evaluated. Although this a very interesting and important question that will be addressed in further analyses and studies, we think it is important to first focus on the process of the self-evaluation. Both the idea of carrying out self-evaluations instead of external evaluations and the idea of evaluating programmes of assessment instead of single assessment methods are relatively new. Future research is still needed here. For example, for formative purposes, further research is needed to explore whether the CAP self-evaluation procedure indeed, as it seems to do, stimulates reflection on CAP quality and if this leads to future improvements of the CAP. With regard to programme quality as opposed to single assessment method quality, there still is a need for more explicit standards specifying acceptable levels of all quality criteria for the programme as a whole. These standards are necessary for summative evaluation of assessment programmes, but can also serve as a point of reference or benchmark for schools when carrying out (formative or summative) self-evaluations.

For now, we conclude that the evaluation of assessment programmes by means of a self-evaluation procedure seems to be possible for formative purposes, but schools need to be supported in the process. A group interview guided by an expert in the field of assessment quality seems necessary to get an overall picture of the CAP's quality. For summative purposes and accountability, though, issues concerning the reliability of self-ratings become more important, and more research is needed on this matter. The combination of formative and summative purposes of self-evaluation, when self-evaluation is used for both school improvement and accountability, could cause problems in this respect. Differences between judges are

generally unwanted in summative evaluation, whereas they may be beneficial for formative purposes by helping generate discussion and stimulate reflection. Finally, with regard to the evaluation of CAP quality, the integral framework of the twelve quality criteria used here seems to be promising to evaluate programme quality in an integrated way.

References

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

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

Baartman, L. K. J., Bastiaens, T. J., Kirschner, P. A., & Van der Vleuten, C. P. M. (2007b). Teachers' opinions on quality criteria for Competency Assessment Programmes. *Teaching and Teacher Education, 23*, 857-867.

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 and 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.

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

Black, P., & Wiliam, D. (1998). Assessment and classroom learning. *Assessment in Education: Principles, Policy & Practice, 5*, 7-75.

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

Clark, D. B., & Sampson, V. D. (2005). Analyzing the quality of argumentation supported by personally-seeded discussions. In T. Koschman, T. Chan & D. D. Suthers (Eds.), *Computer-supported collaborative learning 2005: The next 10 years!* (pp. 76-85). Taipei, Taiwan: Lawrence Erlbaum.

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., Gijbels, D., & Van de Watering, G. (2004, June). *Assessment engineering: Aligning assessment, learning and instruction*. Keynote lecture, EARLI-Northumbria Assessment Conference, Bergen, Norway.

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

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

Glaser, R., & Silver, E. (1994). Assessment, testing and instruction: Retrospect and prospect. *Review of Research in Education, 20*, 393-419.

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

Hickey, D. T., Zuiker, S. J., Taasoobshirazi, G., Schafer, N. J., & Michael, M. A. (2006). Balancing varied assessment functions to attain systemic validity: Three is the magic number. *Studies in Educational Evaluation, 32*, 180-201.

Jonsson, A., Baartman, L. K. J., & Lennung, S. (2007). *Estimating the quality of new modes of assessment. The case of an "Interactive Examination" for Teacher Competency*. Manuscript submitted for publication.

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.

Knight, P. T. (2000). The value of a programme-wide approach to assessment. *Assessment & Evaluation in Higher Education, 25*, 237-251.

Kuhn, D. (1994). *The skills of argument*. Cambridge, UK: Cambridge University Press.

Kyriakides, L., & Campbell, R. J. (2004). School self-evaluation and school improvement: A critique of values and procedures. *Studies in Educational Evaluation, 30*, 23-36.

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.

Maclellan, E. (2004). Initial knowledge states about assessment: novice teachers' conceptualisations. *Teaching and Teacher Education, 20*, 523-535.

McNamara, G., & O'Hara, J. (2005). Internal review and self-evaluation – The chosen route to school improvement in Ireland? *Studies in Educational Evaluation, 31*, 267-282.

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

Miles, M. B., & Huberman, A. M. (2003). *Qualitative data analysis. A sourcebook of new methods* (2nd ed.). Beverly Hills, CA: Sage Publications.

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.

Nevo, D. (1994). Combing internal and external evaluation: A case for school-based evaluation. *Studies in Educational Evaluation, 20*, 87-98.

Nevo, D. (2001). School evaluation: Internal or external? *Studies in Educational Evaluation, 27*, 95-106.

Scheerens, J., Van Amelsvoort, H. W. C. G., & Donoghue, C. (1999). Aspects of the organizational and political context of school evaluation in four European countries. *Studies in Educational Evaluation, 25*, 79-108.

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

Scriven, M. (1991). *Evaluation thesaurus* (4th ed.). Thousand Oaks, CA: Sage Publications.

Simon, S., Erduran, S., & Osborne, J. (2006). Learning to teach argumentation: research and development in the science classroom. *International Journal of Science Education, 28*, 235-260.

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

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.

Toulmin, S. (1958). *The uses of argument*. Cambridge, UK: Cambridge University Press.

Valli, L., & Rennert-Ariev, P. (2002). New standards and assessments? Curriculum transformation in teacher education. *Journal of Curriculum Studies, 34*, 201-225.

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

6. Developing high-quality Competence Assessment Programmes: A cross-case analysis[6]

Abstract

As assessment methods are changing, questions of what constitutes good assessment in competence-based education arise. Therefore, assessment innovations were studied in two contrasting institutions for vocational education in the Netherlands: a 'traditional' and an 'innovative' school, both working towards competence-based education. They were compared with regard to assessment characteristics and quality. A self-evaluation procedure was used with which the schools evaluated their own assessments based on 12 quality criteria for Competence Assessment Programmes (CAPs). Results show that the two schools seem to operate from different frames of reference. They use different assessment methods, and different approaches to assure assessment quality. The innovative school seems to be more aware of its own strengths and weaknesses, seems to have a more positive attitude towards teachers, students, and educational innovations, and explicitly involves stakeholders (i.e., teachers, students, and the work field) in their assessments. This school also had a more explicit vision of the goal of competence-based education and could design its assessments to stimulate these goals.

[6] This chapter is based on:
Baartman, L. K. J., Prins, F. J., Kirschner, P. A., & Van der Vleuten, C. P. M. (submitted). *Developing high-quality Competence Assessment Programmes: A cross-case analysis*.

Introduction

Competence-based education has become popular in many European countries, both at the level of policy-making and at the level of educational practice. In the US, a similar movement towards what is called performance standards-based education can be observed (Valli & Rennert-Ariev, 2002). Competence is seen as the capacity to enact specific combinations of knowledge, skills and attitudes in appropriate job contexts (Lizzio & Wilson, 2004), and as such is expected to reduce the gap between the labour market and education. In vocational education, the context of this study, there is a special recognition of the need for education directed at competence development, and not just at attaining qualifications (Biemans, Nieuwenhuis, Poell, Mulder, & Wesselink, 2004).

Important in these changes is to re-think what adequate assessment methods are (Biggs, 1996; Birenbaum, 1996). Competence-based curricula require specific assessment approaches to adequately determine competence acquisition, not only for assessment alone, but also because of its influence on student learning, teaching practices, and educational development (Alderson & Wall, 1993; Myers & Myers, 2007, Barnes, Clarke, & Stephens, 2000). In this light, traditional knowledge-focused assessments are criticised for being too limited in scope and not being valid for assessment in competence-based education (e.g., Birenbaum et al., 2006; Linn, Baker, & Dunbar, 1991). On the other hand, the reliability of new forms of assessment such as portfolios and work-based assessments should be improved before they are to be used for high-stake purposes (e.g., Baume, Yorke, & Coffey, 2004; Klein, McCaffrey, Stecher, & Koretz, 1995). In previous studies, we argued for Competence Assessment Programmes (CAPs) to assess competence acquisition. CAPs combine traditional tests and recently developed assessment methods and involve both summative and formative assessments (Chapter 2; Baartman, Bastiaens, Kirschner, & Van der Vleuten, 2007a). By applying a programme-wide approach to assessment, problems with regard to developing both valid and reliable assessments could be reduced (Knight, 2000).

The current study focuses on the quality of such CAPs in vocational education in the Netherlands, which is in the middle of the development towards competence-based education. Three recent national developments are of influence on this study. First, as a governmental policy, all vocational institutions are legally bound to adopt a competence-based curriculum from 2010 on. At the moment, Knowledge Centres for VET (Vocational Education and Training) and Industry, which involve representatives from social partners and vocational institutions, are developing new competence-based national qualification profiles. Secondly, from 2004 on, the external monitoring of the quality of examinations in vocational education has been carried out by a single body, the Examination Quality Centre (EQC), which was established by the Ministry of Education to improve assessment quality. Recently, new quality standards have been developed to which all institutions must conform. These standards are formulated at a higher aggregation level than before to allow greater freedom to the institutions to implement innovative competence

assessments. Thirdly, the monitoring system itself is being adapted. Vocational institutions now have to first carry out a self-evaluation of the quality of their assessments, which forms the basis for a more or less extensive external follow-up. This approach is normal at institutions of higher education where an external audit is preceded by a 'self-study' based upon a set of evaluation criteria. The main consequence of these developments is the increased freedom allowed to the institutions to implement different forms of competence-based curricula and assessments, accompanied by a greater responsibility for demonstrating quality.

At the moment, no systematic research has been conducted into how schools are innovating their assessments as a reaction to these developments, and very few examples are available of CAPs that are both valid for the assessment of competence, and reliable enough to make high-stakes decisions. The goal of this study was to compare the CAP characteristics and CAP quality of a more traditional and a more innovative school to examine how innovations are carried out in practice, without pretending to offer clear-cut solutions as to what constitutes an ideal CAP. It needs to be noted here that the terms 'innovative' and 'traditional' are mainly used here to describe the differences between the two schools. Innovative means that this school is actively working towards competence-based education, whereas the traditional school is more reluctant in this respect, which does not mean that this school is used here as a negative example. The traditional and innovative school were selected from eight schools participating in a larger research project, all working towards competence-based education, but differing in their innovations and experiences with competence-based curricula.

Determining assessment quality

To evaluate CAP quality in this study, a self-evaluation procedure was used which enables schools to evaluate the quality of their own CAP (see the methods section). A self-evaluation was chosen because it is becoming an increasingly important approach to both accountability and school improvement, in both the Netherlands and many other countries (McNamara & O'Hara, 2005). A previous study on the use of this procedure showed that schools are able to evaluate their own CAP for formative purposes (i.e., school improvement), but that the procedure may not be reliable enough for summative purposes such as accountability (Chapter 5; Baartman, Prins, Kirschner, & Van der Vleuten, 2007). The self-evaluation was based on 12 quality criteria developed and validated in earlier studies (Chapters 2, 3, and 4; Baartman, Bastiaens, et al., 2006, 2007a, 2007b). Traditional quality criteria like validity and reliability are not fundamentally unsuited for CAPs, but their relevance for assessing competence needs differentiation (Stokking, Van der Schaaf, Jaspers, & Erkens, 2004). To this end, the traditional criteria of reliability and validity were operationalised in a way as to make them more suitable for use in competence-based education (Chapter 2; Baartman, Bastiaens et al., 2007a). They were then complemented with criteria reflecting new ideas about good assessment in competence-based education, such as the meaningfulness, cognitive complexity and consequences of an assessment (e.g., Kane, 1992, 2004; Linn et al., 1991; Van der

Table 6.1. Short description of the twelve quality criteria for CAPs

Criterion	Short description
Fitness for Purpose	Alignment among standards, curriculum, instruction and assessment. The assessment goals and methods used should be compatible with the educational goals
Comparability	CAPs should be set up and carried out in a consistent way. The tasks, criteria and working conditions should be consistent with respect to key features of interest, and scoring should occur in a consistent way
Reproducibility of decisions	The decisions made on the basis of the results of CAP should not depend on the assessor or the specific assessment situation. Therefore, multiple assessors, assessment tasks and situations should be combined
Transparency	CAPs should be clear and understandable to all stakeholders. Learners and assessors should know the scoring criteria, and the purpose of the assessments. External controlling agencies should be able to get a clear picture of the way in which a CAP is developed and carried out.
Acceptability	All stakeholders should approve of the assessment criteria and the way the CAP is carried out. They should have confidence in the CAP's quality
Fairness	Students should get a fair chance to demonstrate their competences, for example by letting them express themselves in different ways and making sure the assessors do not show biases
Fitness for Self-Assessment	CAPs should stimulate self-regulated learning. They should include specific methods to foster this learning such as practice in self-assessment and giving and receiving feedback
Meaningfulness	CAPs should have a significant value for all stakeholders involved. For learners, assessments should be a learning experience in themselves, and be useful for the learning process. For teachers and employers, the assessments should be meaningful in terms of the requirements of the future job
Authenticity	The degree of resemblance of a CAP to the future workplace, in terms of the assessment task, the physical and social context, and the assessment criteria
Cognitive complexity	A CAP should reflect the presence of the cognitive skills needed and should enable the judgment of thinking processes
Educational consequences	The degree to which the CAP yields positive effects on learning and instruction, and the degree to which negative effects are minimised
Costs and efficiency	The feasibility of developing and carrying out the CAP for both students and assessors, and the time and resources needed, compared to the benefits

Vleuten & Schuwirth, 2005). This resulted in the 12 quality criteria for CAPs used here (see Table 6.1).

To study innovations towards assessment of competence in the two vocational schools, three research questions were formulated that guided our analysis: (1) How do the traditional and the innovative school use the 12 quality criteria to evaluate their CAP? Do they use the same or different approaches to assure CAP quality? (2) Does the innovative school's CAP better comply with new, competence-based quality criteria (i.e., *acceptability*, *authenticity*, *cognitive complexity*, *educational consequences*, *self-assessment*, and *costs and efficiency*)? and (3) How could the differences between the two schools be explained?

Method

Participating schools

The two schools participating in this study were selected from eight vocational schools participating in a larger research project. All schools offer laboratory technology education, a vocational programme preparing students for a job as a laboratory assistant or a laboratory technician. They were part of a national consortium aiming at the innovation of technical education, which started in 2000 with the development of problem-based lesson materials called 'unit books' and has now started with the development of competence-based materials called 'project books'. Consortium schools are free to use the unit books or project books, and to modify the pedagogy and assessments to their specific needs. The main changes in the competence-based project books compared to the problem-based unit books are an increased emphasis on the importance of assessment in the professional job context, a greater emphasis on attitudes, and a less prescriptive character in order to stimulate students to regulate their own learning (Klatter, 2006). For this study, two contrasting cases (Yin, 2002) were selected based on their CAP characteristics. An overview was made of the CAP characteristics of all eight schools of the larger project (see Table 6.2 for a concise version). Five schools work with the problem-based unit books, and three work with the competence-based project books. Furthermore, the schools differ with regard to the assessment methods they use. Based on Table 6.2, two schools were selected: 'traditional' school C and 'innovative' school H. School C had only recently started to work with the unit books and the participants from this school were relative novices in the use of new assessment methods. For our comparison, this school was preferred above school B, which has similar CAP characteristics, but of which less documentation was available. School H works with the project books and was developing and pilot testing an entirely new CAP at the time of data collection. It needs to be noted that school C and school H evaluated CAPs of different school years, which may make them less comparable in this respect. This was unavoidable because the unit books are only used in school years 1 and 2, while the project books are only used in year 3.

Table 6.2. Summary of the CAP characteristics of eight schools participating in the larger research project. Schools C and H were selected for this in-depth comparison

	Start problem-based education	School year	Didactical method	MC test	Written test – open questions	Assessment of products made	Assessment interview	peer/self assessment	Observation in simulated situation	Presentation	Criterion-based interview	Observation in the workplace	Proof of competence	Portfolio
School A	2003	1st & 2nd	unit books	x	x	x	x	x	x			x	x	
School B	2003	1st & 2nd	unit books	x	x	x	x	x	x					
School C	2005	1st & 2nd	unit books	x	x	x	x	x	x					
School D	2004	1st to 4th	unit books	x	x	x	x	x	x	x		x		
School E	2003	1st to 4th	unit books	x	x	x	x	x	x	x		x		
School F	2003	3rd	project books		x	x	x	x	x	x	x			
School G	<2003	3rd	project books		x	x	x	x	x	x	x			
School H	2003	3rd	project books			x	x	x		x	x	x		

CAP quality self-evaluation procedure

A short description of the self-evaluation procedure is given here. For a more elaborate description and an evaluation of the self-evaluation process itself, see Chapter 5, or Baartman, Prins et al. (2007). In this study, three functionaries from each school participated in the self-evaluation: the department manager, a member of the examination board and another teacher. An earlier pilot study showed that these three functionaries generally are well-acquainted with the assessments used. Together, they have a complete overview of all assessments used, in terms of national and school-specific policies and regulations, and from personal practical experience. The self-evaluation procedure consisted of two phases. In the first phase, participants individually evaluated their CAP by means of a web-based tool. They were asked to indicate the assessment methods in their CAP (those in Table 6.2 were given as options, and others could be added) and to evaluate this CAP on the 12 quality criteria. All quality criteria had been further operationalised into indicators, providing concrete quality aspects observable in practice (see Table 5.2 on page 101, and Baartman et al., 2007). For each indicator, a quantitative and a qualitative judgment were given. Quantitatively, participants moved an analog slide-bar from 'not at all' to 'completely'. A 'don't know' option was also available. Behind this slide bar was a rating scale from 0 to 100, which was invisible as not to give the idea of giving a score or mark. Qualitatively, the participants supported each rating by an example from their own CAP. The second phase consisted of a group interview. All individual input from the first phase was assembled in an overview of CAP quality, which formed the basis for the group interview in which the different ratings and examples were discussed. The group interview lasted approximately two hours and had a semi-structured character. First, the group was asked to globally describe their CAP. This was followed by a discussion in which the participants were explicitly encouraged to comment on their own and each others' ratings and examples. The interviewer asked for further information or explanation if the argumentation was unclear to the interviewer or the evaluators had clearly different opinions, as apparent from the overview. Finally, besides evaluating their CAP, participants were asked to provide documentation of their assessments such as policy documents, overviews of assessment procedures and criteria, and guidelines for students and teachers.

Data analysis

The available data sources were the web-based self-evaluations, the group interviews and the additional documentation provided by each school. All group interviews were transcribed verbatim. Only the qualitative data of the self-evaluations were used, as our goal was to explore the innovation of CAPs in competence-based education, and not so much the rating participants attributed to their CAP. The two schools were systematically compared using the 12 criteria as a conceptual framework for the analyses. Miles and Huberman's (2003) method of

cross-case comparison was used, where qualitative data are first meaningfully reduced or reconfigured (data reduction), then organised into different displays such as diagrams or matrices (data display), from which conclusions are drawn and verified in the last phase (conclusion and verification). For the purposes of this research, the findings were not fully developed as individual, complete and descriptive cases, but rather as interpretive comparisons on the 12 criteria. To answer the first two research questions on the use of the 12 quality criteria and differences between the CAPs, a summarising display was constructed for each case, providing the schools' examples given of the 12 criteria, summarised over the three participants. A description of the two CAPs was added to study the relationship between CAP characteristics and CAP quality as reported in the self-evaluation. A check (verification) was carried out by an independent researcher not involved in the current project, who independently reconstructed the displays. Only very small differences between the two researchers were found, which were discussed and changed in accordance with both researchers' opinions. The displays of the two cases were then assembled in a meta-matrix which enabled the systematic comparison of the two cases on each of the 12 quality criteria. Finally, possible factors influencing the differences between the two cases (research question 3) were first identified by the first author during the other analyses, and noted down as hypotheses about general similarities and differences between the two schools (e.g., the innovative school involves stakeholders, whereas the traditional school does not). The first and second author then independently re-analysed the data displays, going back to the original interviews when necessary, looking for evidence and counter-evidence of the hypotheses. The findings of these two independent analyses confirmed all hypotheses except one (i.e., that both schools often refer to the national consortium to account for the quality of their CAP; the traditional school appeared to do this much more often than the innovative school). All confirmed hypotheses are presented in the results section.

Results

Before the results of the three research questions are presented, the CAP characteristics of the traditional school C and the innovative school H are further described in the next two sections.

Traditional school (C): CAP characteristics

The traditional school only recently started to implement problem-based education, in the school year 2005-2006. Their assessment programme consists of four parts, which are typical for the schools working with the problem-based unit books (see also Table 6.2). In general, the students use the unit books to work on 15 to 20 practical tasks per term (about 6-8 weeks), such as preparing a lab report or a graph with results. Theoretical knowledge is assessed through an integral theoretical test, taken at the end of each term, consisting of both multiple choice questions and open

questions and its content is connected to the knowledge needed for working on the tasks. The school tried to organise the integral theoretical test around a common theme, for which all subject teachers had to develop questions relevant to their subject, but they encountered some problems with this approach: 'The questions for the integral test are provided by the different subject teachers. We tried to link all questions to a common theme (…) but when I hear the discussions and stories about it, you see it doesn't work. Some people even suggested just cutting out the theme, they think it is nonsense. So in my opinion it is not really an integral test, it is a combination of different subjects' [school C]. Second, the products made while working on the tasks are assessed by a teacher and have to be judged as satisfactory, and three tasks per period are selected for a more thorough summative assessment (i.e., assessment of products made). The mean grade for these tasks forms the test result. Third, an assessment interview is taken at the end of each term. A number of aspects are selected on which the students are assessed during that term, for example their attitude towards the learning process, functioning in the group, and study skills. As input for the assessment interview, all students assess themselves and their peers on an assessment form, as do the teachers. All input is discussed during the interview, with the teacher making the final decision and setting the learning goals for the next term. Finally, practical skills (e.g., preparing a microscope slide) are assessed by the teacher during the entire term while working in the laboratory at school (i.e., observation in simulated situation). A list of practical skills is constructed for each term, and students have to demonstrate all skills to a teacher, who ticks off the skill on the list when it is assessed as satisfactory.

Innovative school (H): CAP characteristics

The participants of the innovative school started to work with the unit books in 2003 and were developing an entirely new CAP at the time of data collection. While the participants of some schools had difficulties exactly describing their CAP during the group interview, the entire teacher team and employers were involved in the development of the new CAP in this school and thus all participants had a clear overview of their CAP. The new CAP was still under construction, and although parts of it were pre-tested with students and internship supervisors, no actual user experiences were available. The main part of all assessments is carried out in the workplace during internships and no separate multiple choice tests or other knowledge tests are used. First, the assessment of products made involves the tasks in the project books, which are carried out by all students during their individual internships. All students work in a company for four days a week, and come back to school one day a week to discuss the project tasks, the planning of the projects and to study the theoretical knowledge necessary to perform the project tasks. The students work in small project teams; during the school day the different individual tasks are combined into one large group task. Second, an assessment interview is used to assess students' functioning in the project team. The students fill out an assessment form for themselves and their group members, which are discussed during an assessment interview with the teacher where learning goals are set for the next

project. Third, at the end of the term, the project teams give a presentation of their project, to which a criterion-based interview is connected. The presentation is given as a group, and questions are asked to each individual team member to assess the individual contributions. Questions are asked by teachers, internship supervisors and students, and focus both on the general theory behind the project tasks and the different individual internships. Students are, for example, asked to explain why they carried out the task in that specific way in their company. Finally, an important part of the CAP consists of observations in the workplace, mainly carried out by the internship supervisor. To facilitate and guide these assessments, the school uses an overview of all competences to be developed by the students during the course of the educational programme, for example carrying out different kinds of analyses, and communicating with clients. The competences are worked out in different phases of development, with the first phases describing for example easy analyses carried out under supervision, while in later phases the student has to understand why certain analyses are carried out in each situation, and has to show a critical attitude towards his or her own work. The overview of competences was developed by the school in co-operation with the national knowledge centre for Vocational Education and Training (VET) and Industry, and the regional work field. The internship supervisors and the students themselves use the overview to assess the student's work in the company. The overview is discussed with the teacher and the student sets specific learning goals for the next term.

General similarities and differences between the schools

Now the CAP characteristics of both schools have been described, the next step is to look into the quality of these CAPs. Because two extreme cases were chosen for this comparison, it is not surprising that the CAPs differ on almost all quality criteria. Still, some general similarities and differences were found, which are not specific to one or two quality criteria, but run throughout the entire comparison. They are summarised in Table 6.3. First, the participants from both schools were willing to be self-critical. They reported many different problems with regard to their assessments and were willing to discuss the advantages and disadvantages of their approach. However, they also often emphasised they are still in a developmental phase towards competence-based education, and improvements are continuously being designed and implemented. Both schools reported that a 'culture shift' towards competence-based assessment takes a long time. The third general similarity pertains to the many references to the EQC. Both the national standards and the self-evaluation are new to vocational education, and both schools struggled with their new responsibility to account for the quality of the examinations to the EQC.

With regard to the general differences, the schools seem to judge their CAPs from different frames of reference. First, their attitude towards students, teachers, the work field, and educational innovations as a whole is very different. While the innovative school is quite positive in this respect, the traditional school mentions many problems with for example teachers and students who still have to get used to

Table 6.3. General similarities and differences between the traditional school and the innovative school

General similarities	
Description	**Example**
Both schools are self-critical towards CAP quality	'A lot has to change before we have real competence-based education'
Both schools are still in the middle of the development process towards competence-based education	'The assessments are still under construction. That will take another few years'
Both schools often refer to the Examination Quality Centre (EQC) and how they have to account for the quality of their examinations	'The choice for the summative assessments also depends on the prices of the EQC'

General differences		
Description	**Example innovative school**	**Example traditional school**
The innovative school has a more positive attitude towards students, teachers and innovations in general	'Students formulate their own learning goals'	'If students get feedback, they do not know what to do with it. They cannot regulate their own learning'
The innovative school is more pro-active: when they encounter problems, they mention concrete possibilities for improvement. The traditional school does not	'We need to further specify how we want students to function in the workplace. Internship supervisors need to be trained'	'Teachers and students experience problems with the integrated assessment' (no improvements)
The innovative school has a more explicit vision of competence-based education	'Our goals is to deliver competent professionals, therefore we assess in the workplace'	'We do not have a clear picture in mind of the learning goals of competence-based education'
The innovative school explicitly involves stakeholders in their assessments. The traditional school does not	'We discussed the assessments with the teachers and internship supervisors, and we piloted it with the students'	'We never explicitly measured or asked this'

a competence-based learning environment. Second, when the schools do encounter problems, they react in a different way. The innovative school mentions possible solutions and concrete plans for improving almost all problems they encountered during the pilot phase of their new CAP, while the traditional school is not sure what to do with the problems they encounter. This is also related to the third general difference, namely that the innovative school has a clear vision of competence-based education and what they want to achieve with their new CAP. The traditional school, on the other hand, does not yet have a clear picture in mind of the characteristics and goals of competence-based education, which makes it difficult to make more concrete plans for improvement. Finally, the innovative school is much better informed about their stakeholders' opinions about their CAP, and explicitly involved the stakeholders in the development process, while the traditional school implemented the unit books and associated assessment methods developed by the consortium, without for example taking into account teachers' opinions, who were afraid their workload would increase.

In the next sections, the more specific similarities and differences between the traditional and the innovative school for each of the 12 quality criteria are discussed, and examples are given to underline these differences. Also, the general results described here are highlighted again.

Fitness for purpose

Fitness for purpose is a basic quality criterion for CAPs as it relates the goals of education to the goals of the assessment and prescribes that the two of them must be well-aligned (Brown, 2004; Miller & Linn, 2000). Therefore, the development of the new national qualifications into competence profiles is very important here, as the schools have to base their assessments on these profiles and have to prove to the EQC that their assessments adequately cover them. School C worked with the older non-competence-based national qualification structure, but encountered difficulties relating its problem-based assessments to these qualifications, whereas the new competence-based ones were not available yet: 'It is very difficult to relate our assessments to the qualifications … that doesn't work anymore, and we cannot refer to the new competence profiles' [school C]. Moreover, the participants did not have a clear picture of the new competence profiles to guide the development of new assessments: 'I think we do not yet have a clear picture in mind of the actual learning goals of this new type of education (…) the only thing I know is that it is very different from what we have done so far' [school C]. School H, on the other hand, co-operated with the national knowledge centre and the work field to develop a new competence profile for laboratory sciences. The competence overview the school uses as a basis for the assessment in the workplace 'does not separately describe knowledge, skills and attitudes' [school H], and thereby stimulated the integrated assessment of competence. School C on the other hand struggled with the integrated assessment of knowledge, skills and attitudes. Although it referred to its integrated knowledge test, integration in this test only means that knowledge questions are asked about the same themes that also formed the basis for the tasks in the unit

books, but knowledge, skills and attitudes are not actually assessed in an integrated way in a work situation. School H seems to be further on the way towards integrated assessment, as it actually integrates knowledge questions into the criterion-based interviews about the work carried out in the workplace. This school considers knowledge to be conditional for performance, but some doubts may arise as to whether the required knowledge base can be adequately assessed through a criterion-based interview alone (Valli & Rennert-Ariev, 2002). In sum, the new competence profiles seem to offer better opportunities for integrated assessment of competence, but the question of how to assure adequate assessment of a necessary knowledge base at the same time needs further investigation.

Comparability

The second criterion for CAPs is *comparability*, which is related to reliability (Chapter 2; Baartman, Bastiaens, et al., 2007a) as it was used for traditional assessments. Both the traditional and the innovate school deem comparability very important. This is interesting, because comparability is more difficult to achieve in competence-based education as less standardised tasks are used. Comparability, therefore, is worked out in different ways in both schools, as became apparent from the examples the schools gave for this criterion. Traditional school C administers the same knowledge tests to all students at the same time, and uses strict scoring rules for the assessment of skills in the laboratory class. Explaining why they think their CAP is comparable, the participants mainly referred to standardisation of tasks, conditions, criteria and procedures: 'We pay a lot of attention to comparability, to get everything as objective as possible. All procedures are laid down, all tests and criteria are put together in a matrix. I think everything is perfect in this respect' [school C]. The innovative school H could not refer to standardisation, because its students are assessed in different companies. It did, however, take comparability into account: 'We do make a difference between companies … in some, students can perform routine tasks, but not the more advanced project books in which they have to experience the entire complexity of laboratory work' [school H]. Interestingly, the participants also referred to the criterion *reproducibility* as a way of ensuring reliable assessments without necessitating full comparability: 'The procedures are comparable, but you can never prevent small differences between companies. The only way to justify these differences is to assess multiple internships in different companies' [school H]. This comparison shows that *comparability* can be achieved in many different ways, and that CAPs in competence-based education can be adequately comparable if some measures are taken to assure comparability, without necessitating full standardisation (Benett, 1993).

Reproducibility of decisions

Reproducibility of decisions, which was already shortly referred to, is another quality criterion related to the traditional idea of reliability. By using multiple assessments, carried out by multiple different assessors, a reliable picture of a student's competences can be obtained, without necessitating standardisation (Moss, 1994;

Van der Vleuten & Schuwirth, 2005). It became apparent that traditional school C mainly tries to ensure reliable assessments by standardisation and objectivity. It is therefore not surprising that this school focuses less on reproducibility as a way of achieving reliability, whereas the innovative school does. In school C, most of the time only one assessor, the teacher, is involved: 'The integrated knowledge tests are constructed and assessed by multiple assessors, but each assessor only assesses one part of the test (…) I think it also depends on the assessment method, if you need multiple assessors. When you use a written test with a clear answer specification and clear standards, you need only one assessor, but if you assess the student's functioning in a job situation, multiple opinions generate a more complete picture' [school C]. In contrast, for school H reproducibility is the main way of achieving reliable assessments, as it could not and did not want to make its assessments entirely comparable. It uses the competence overview to assess students during the work on the project books and to monitor competence development (multiple times), and it involves multiple different assessors (teachers, students, internship supervisors) in the assessment interview, the presentations, and the criterion-based interviews. Looking at the different approaches to *comparability* and *reproducibility* taken by the two schools, *comparability* seems to be a more traditional way of achieving reliable assessments, whereas *reproducibility* offers a different way of ensuring reliability that is more suitable for competence-based education. It is important to note that *reproducibility* can be achieved without full *comparability*.

Transparency

The criterion *transparency* prescribes that a CAP should be clear and understandable to all stakeholders involved in the assessments (Frederiksen & Collins, 1989; Linn et al., 1991). Here, transparency relates to students, teachers and the work field as stakeholders. Both schools are satisfied with the transparency of their assessment, because procedures and criteria are specified. It is not common practice, however, to actually check whether the stakeholders understand what is specified. School H carried out some checks, but it also acknowledged that its CAP is completely new: 'I think it is not clear to everybody yet. I mean, you can put things on paper, but it is one step further to actually understand it and to grasp the meaning of it. So everything has been laid down, but if our students and the supervisors really understand the assessment, and the consequences of these new assessments? I think that will take another few years' [school H]. The two schools seem to assure transparency towards teachers and students in different ways. With regard to the teachers, both schools reported that the assessments are discussed in the team, but they also acknowledged that their teachers are more familiar with traditional testing than with newer forms of assessment like assessment interviews. In contrast, procedures and criteria are not usually discussed with the students. The traditional school assumed its students to understand the assessments, because they are instructed to carefully study the guidelines and do not ask many questions about the assessments. School H referred to the trials in which students gave positive reactions, and the fact that the students themselves have to fill out the competence

overview and thus have to understand it in order to be able to assess themselves. This difference again seems to show that the innovative school worked out the criteria differently than the traditional school. It was not satisfied with 'just' laying down criteria and procedures, but was aware of the fact that the stakeholders have to understand the CAP before being able to adequately work with it.

Acceptability

Acceptability adds to the *transparency* criterion that stakeholders should approve of the procedures and criteria used, and have confidence in the quality of the CAP (Stokking et al., 2004). The most salient difference between traditional school C and innovative school H seems to be that the stakeholders in school H are actually involved in the assessments, whereas school C did not ask their opinion. The fact that school H completely built up its CAP from scratch seemed to be an advantage, as all stakeholders could be involved from the beginning of the development process, which seemed to have increased acceptability: 'I tried this out with a couple of students, and I asked them, can you work with this and do you have any questions … they thought we did not ask any strange things, they agreed with the criteria (…) and the teachers, we all agree about it, the new assessment is an improvement (…) and the work field, the people I talked to, they thought it is more concrete, they are forced to look more carefully at how the student is working during internship, and not just saying, oh I think it is OK' [school H]. A second interesting result is that during the group interview, the participants from school C became aware of the fact that they did not know their stakeholders' opinions: 'The integrative knowledge test causes problems; maybe this is because the students got too little feedback during their learning process. But there could be many more causes … that's the idea of this evaluation, isn't it, to get clear where your CAP needs improvements (…) you cannot say just out of the blue what students think of the assessments. We should ask them more specifically, interview them, or give them a questionnaire' [school C].

Fairness

Fairness comprises a number of different indicators relating to procedures to rectify any mistakes made during the assessment, the use of various assessment tasks, the fact that assessors should not be prejudiced and finally that students should perceive the assessments to be fair (Dierick & Dochy, 2001, Hambleton, 1996; Linn et al., 1991), as this perception can have huge influences on for example *acceptability*. The results seem to show the same pattern seen for acceptability. The traditional school C did not investigate whether assessors are prejudiced or not, and whether its students think the assessment is fair, but it assumed it to be so: 'I take it for granted our teachers are not prejudiced, they have a professional attitude' and 'we did not ask the students specifically, but complaints about unfair assessments are very rare' [school C]. On the other hand, school H seems to be more aware of the fact that it cannot just assume its CAP to be fair: 'As far as I can say, our assessors are not prejudiced, but that's my personal opinion. My experiences are based on the small

sample with which I piloted the assessment' [school H]. A tentative conclusion that seems warranted here is that both schools do not yet have adequate solutions to solve all fairness issues, but the innovative school seems to be more aware of the measures it has to take to assure *fairness*, whereas the traditional school takes it for granted that its assessments are fair.

Fitness for self-assessment

Fitness for self-assessment prescribes that a CAP should not only enable the judgement of a learner's competences, but also stimulate self-regulated learning, a learning goal that has become more prominent with the development of competence-based education (Tillema, Kessels, & Meijers, 2000). It is therefore not surprising that school H pays much more attention to this quality criterion than school C. The participants of school H describe how their CAP stimulates self-regulated learning: 'Yes, our students assess themselves and each other when they fill out the competence overview … and based on that we have the assessment interview, in which they get feedback, and they have to say themselves what they want to improve and work on in the next term' [school H]. School C, on the other hand, did consider self-regulated learning very important, but its CAP failed to stimulate this. Main problems were that almost no feedback was given on the tasks in the unit books, that teachers felt resistance towards giving this feedback because it increased their workload, and that giving feedback in itself was very new to teachers, who therefore could not communicate to students the importance of feedback. One teacher remarked: 'I have little experience with that, but I notice that students do something with my feedback. They think it is positive they get feedback and try to improve their work. But only on technical matters, for example how do you tackle this problem and which method do you use here … but things like what kind of person am I, functioning in the group, how do I behave towards other students … that is very difficult. As a teacher, I know the technical part much better' [school C]. Concluding, *fitness for self-assessment* seems not only to depend on the design of the CAP itself, but also on the way teachers or assessors actually carry out the CAP in practice.

Meaningfulness

Like *fitness for self-assessment*, *meaningfulness* is a quality criterion that is closely related to the ideas of good assessment in competence-based education. It describes how assessments should be meaningful learning events in themselves, for example by the feedback they generate (Linn et al., 1991). In this respect, both schools seem not yet confident that their CAP is meaningful in the eyes of students, teachers and internship supervisors. School H has only just started the implementation of their new CAP, and acknowledges that an evaluation of meaningfulness is necessary some time after the new CAP has been fully implemented. School C also signals some problems: in general, there are too few opportunities to get feedback, but a second problem is that students do not use the opportunities they are offered, because they do not recognise assessments as opportunities to learn: 'It does not

come naturally, the assessment system has to encourage students to use the feedback opportunities they get. We have special extra practice sessions for that. But it is not easy, because if you say, come to me if you have any questions, than suddenly they don't have any questions' [school C]. A second interesting result is that teachers and employers do not always perceive new assessments to have an added value. They seem to be afraid the knowledge level of students is not adequately assessed. Concluding, *meaningfulness* still seems to be a difficult quality criterion, and teachers seem to express some resistance towards new assessment methods.

Authenticity

Authenticity relates to the resemblance of the CAP to the future job situation in terms of assessment tasks, working conditions, social context and criteria (Gulikers, Bastiaens, & Kirschner, 2004). Both schools seem to be quite satisfied with the authenticity of their CAP, but an interesting difference between the two schools is that school C often refers to the unit books and the work of the national consortium to account for the authenticity of their assessments, whereas school H refers to the fact that their assessments are carried out in the actual workplace. This illustrates that the schools appear to have different frames of reference from which they judge their CAP. School C only recently started to work with the unit books, in which the tasks are more explicitly related to the job context than in the assessments they used before: 'I think that is the strength of the unit books, all tasks are related to the job situation in some way. Although it is not in the real job environment, it is still recognisable for the students' [school C]. School H has worked with the unit books for a number of years, and is not satisfied anymore with 'just' relating tasks to the job context, but takes a next step towards authenticity by assessing in the workplace itself: 'I think that is the strongest aspect of our new educational concept, the fact that students are actually assessed in the work place' [school H]. Secondly, differences between the schools could be caused by the fact that school C evaluated the CAP of its first and second school year, whereas school H evaluated the CAP of its third year. In their first year, students still have to learn to master the basic skills of laboratory work, whereas the tasks become more complex in successive years.

Cognitive complexity

Like *authenticity*, *cognitive complexity* relates a CAP to the future work situation, but it pertains to the thinking processes a competent professional uses to solve problems encountered on the job (McLellan, 2004). Assessment should, thus, not only focus on the product, but also on the thinking processes: how and why did students act and make choices during their work on a task. The results show that this quality criterion is still quite new to both schools, though they deem it important and plan to pay more attention to it. Both schools referred to the unit books and project books developed by the national consortium in co-operation with the work field, and explained that the completion of these tasks requires the necessary thinking processes. In both schools, though, thinking processes were not explicitly assessed. One participant of school H said: 'I think the thinking processes should be more

explicitly assessed during the presentations and the criterion-based interview. We did not develop that yet, we do not actually ask them how they tackled a problem … but I think you can do it in a criterion-based interview' [school H]. School C thought that the idea of assessing thinking processes better fits in a competence-based approach than in a more traditional learning environment: 'If you take thinking processes into account, and I think that is a really competence-based approach, you need a very open task, for example you give them a substance and they have to find out what it is … and then you assess, how they go about, how they solve this problem' [school C]. These results seem to show that, though the assessment of both product and process is one of the cornerstones of assessment in competence-based education, it is not clear yet how thinking processes could be assessed in practice.

Educational consequences

Educational consequences pertain to the effects of an assessment on learning and instruction, and the curriculum as a whole (Messick, 1994; Schuwirth & Van der Vleuten, 2004). As assessment can have a huge influence, a CAP should stimulate the desired learning processes, and positively influence teachers and students to engage in these learning processes. With regard to this quality criterion, school H is much more positive than school C. School H has a clear view of the desired learning processes, and explicitly tries to design its new CAP in such a way that these learning processes are stimulated. School C on the other hand, does not have a clear picture of the desired learning processes in competence-based education, although it is actively evaluating its CAP. A few remarks made during the interview highlight these differences: 'At the moment, I don't notice any effects of the assessments, like 'I got a bad grade, so now I have to work harder' … but this is also because we ourselves are still struggling with what the desired learning processes actually are' [school C]. And: 'So I am negative about these learning processes, but we are very busy evaluating our assessments at the moment (…) all teachers have come together, because they noticed it was going the wrong way, and we organised some kind of evaluation meeting' [school C]. School H on the other hand described the effects of their CAP like this: 'There is much more direct contact between the internship supervisors and the teachers, and therefore we have a better notion of the knowledge and skills of our students. Teachers now experience teaching as a team task … and you notice that what we teach is much better harmonised with what is necessary on the job. For example, teaching math just for the math is something we don't do anymore' [school H].

Costs and efficiency

Finally, the criterion *costs and efficiency* relates to the feasibility of carrying out the CAP (Hambleton, 1996; Linn et al., 1991). Again, a clear difference can be observed between the schools. Whereas the traditional school reports many problems, the innovative school has explicitly paid attention to feasibility in the design process of its new CAP. Again, it seems to be an advantage that their CAP was designed from scratch, involving the different stakeholders and taking their wishes and possibilities

into account. The participants describe how they designed their CAP: 'We discussed how many days the internship should be … well, to make it more cost-effective we decided for four days internship and one day at school. It has to be attractive to the companies as well, so we gave them two-and-a-half days in which they can determine what work they want students to do. The other one-and-a-half day they work on the project tasks. And we had to cut down the number of theoretical lessons, but you notice that because of the internships and the presentations about theoretical problems, their knowledge is profound enough' [school H]. School C only has a very rough idea of the time and money needed for its CAP, and problems reported include the fact that teachers oppose giving feedback more regularly because they think it takes too much time. When the participants were asked if they think the investments in the CAP outweigh the positive effects, they reacted: 'At the moment there is an atmosphere of disappointment about the effects of problem-based education, at least about the way in which we carry it out now (…) I think, if you can give assessment a function in the educational process, apart from summative testing … if it also generates feedback and guides student development, then I think the effects may outweigh the time it requires. But not if it is only used for summative examination' [school C].

Conclusion and discussion

The goal of this study was to compare a more traditional school and a more innovative school in order to explore how schools innovate their assessments in the transition towards competence-based education. Two contrasting cases were selected, a traditional school and an innovative school, and their Competence Assessment Programmes were described and differences in CAP quality were systematically compared.

With regard to CAP characteristics, the case descriptions showed that the two schools use different assessment methods. In general, the CAP of the innovative school could be regarded as 'more competence-based', based upon observation in the workplace together with presentations and criterion-based interviews. It is remarkable, though, that neither of the two schools use a proof of competence or a portfolio, two assessment methods generally regarded as new assessment methods suitable for competence-based education (e.g., Birenbaum, 1996; Dierick & Dochy, 2001). It needs to be noted, however, that the eight institutions participating in the larger project were discussing the possibility of collaboratively developing a portfolio to be used by all laboratory technology institutions. As both schools participating in this study were part of a national consortium and in that sense were more or less actively working towards competence-based education, this finding may imply that vocational education is indeed still in the middle of the development process towards competence-based education and assessment, and that other schools might be less far in the development towards competence-based education. It also shows, however, that a competence-based CAP, as in the innovative school, does not impertinently have to include a portfolio or a proof of competence.

Another CAP characteristic of the innovative school warranting discussion is the fact that it does not use any separate knowledge tests. Instead, knowledge is assessed through the work on the project tasks, in which knowledge is assumed to be conditional for performance, and through asking questions in a criterion-based interview. Some authors point to the dangers of this development, and warn that assessment of competence should not mean not assessing students' knowledge base at all. For example, Valli and Rennert-Ariev (2002) write that assessments tend to lean 'too much in the direction of craft knowledge to the exclusion of other forms and sources of knowledge' (p. 215), that is theoretical knowledge. Also, Wolf (1989) states that it is dangerous to infer too much from the observation of performance, and that knowledge inevitably needs to be tested independently of performance since this is the best basis for inference beyond the actual situation. Interestingly, Wolf also points to the variety of contexts in which professionals can show their competence. It is exactly this context and task-specificity of performance that makes it difficult to reliably assess in the workplace, as was shown by generalisability studies (e.g., Wass et al., 2001). The innovative school uses its criterion-based interview to make assessment less context-specific, as students are asked to explain why they acted like they did in a specific situation, and how they would do otherwise in another situation. The use of criterion-based interviews might thus be a step towards integrated assessment of competence, taking into account the specific context, but also looking beyond it. More research is needed, though, to investigate if assessment programmes like the one used by our innovative school effectively cover students' knowledge base.

Looking at the quality of the two CAPs, differences were found for almost all quality criteria. A few salient results are highlighted here. First, the innovative school explicitly checked whether its assessment was *transparent*, *acceptable* and *fair* in the eyes of its stakeholders. The participants of the traditional school, on the other hand, merely assumed their stakeholders to be satisfied as they expressed no complaints. Second, the innovative school explicitly designed its CAP to be *fit for purpose*, *fit for self-assessment*, and to generate positive *educational consequences*, whereas the traditional school did not have clear picture in mind of the goals of competence-based education, and thus could not design its CAP to stimulate these goals. Third, the differences between *comparability* and *reproducibility of decisions* show that the innovative school chose to assure reliability through repeated measures by different assessors and in different contexts, while the traditional school focused on standardising assessments. On the other hand, some similarities were found as well. Although the innovative school's CAP was well thought-out, the actual effects on students' learning processes still need confirmation. This shows that the innovative school was also still in the development process towards competence-based education and more evidence is needed of the effects of new assessment methods. Moreover, two other quality criteria caused problems in both schools: *cognitive complexity* and *meaningfulness*. Both schools considered these quality criteria to be important, but they could not give any examples showing that their CAP complied with these criteria.

Concluding, the two schools indeed seemed to work out the criteria in different ways, in which the innovative school seemed to be more aware of its own strengths and weaknesses. The innovative school's CAP seems to better comply with the newer quality criteria *acceptability*, *fitness for self-assessment*, *educational consequences*, *authenticity* and *costs and efficiency*. No differences were found for *cognitive complexity* and *meaningfulness*, as both schools still had problems finding examples of these criteria in their assessments. This result also implies that the quality criteria that still need most attention in both research and practical implementation seem to be the newer ones that came up during the transition towards competence-based education.

Finally, this chapter started with some developments that triggered Dutch institutions to innovate their assessments: the new competence profiles, the new quality standards of the EQC, and the new evaluation procedure carried out by the EQC. As Gonczi (1994) remarked, the development of new assessment methods can never be looked at without taking into account the political and policy perspective. The traditional and innovative school seemed to react in slightly different ways here. The innovative school operates on the forefront of innovation, taking part in an experiment in which the new competence profiles are being developed. This seemed to have offered a head start in the development of new assessments, as adequate knowledge about the new requirements is essential for guiding developments. With regard to the new standards used by the EQC, the innovative school again took part in a trial in which they were evaluated on these new standards, which seemed to offer more freedom to develop a new CAP. As the new standards are more broadly defined, though, it is indispensable that schools are knowledgeable of their assessments, and aware of their strengths and weaknesses. The innovative school seemed to function well in this respect, at least partly because it was developing its new CAP itself and was taking part in the ministry's experiments. This school also seems able to carry out self-evaluations, and did so as a trial (see also Chapter 5, and Baartman, Prins, et al., 2007). Possibly, the fact that schools are required to carry out a self-evaluation in the new evaluation system will cause problems in schools that are less experienced in this respect. The time and effort the innovative school has invested in its developments seems to be a sign here.

References

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

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*, 153-177.

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

Baartman, L. K. J., Bastiaens, T. J., Kirschner, P. A., & Van der Vleuten, C. P. M. (2007b). Teachers' opinions on quality criteria for Competency Assessment Programmes. *Teaching and Teacher Education, 23*, 857-867.

Baartman, L. K. J., Prins, F. P., Kirschner, P. A., & Van der Vleuten, C. P. M. (2007). Determining the quality of competence assessment programs: A self-evaluation procedure. *Studies in Educational Evaluation, 33*, 258-281.

Barnes, M., Clarke, D., & Stephens, M. (2000). Assessment: the engine of systematic curricular reform? *Journal of Curriculum Studies, 32*, 623-650.

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 and Evaluation in Higher Education, 29*, 451-477.

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

Biemans, H., Nieuwenhuis, L., Poell, R., Mulder, M., & Wesselink, R. (2004). Competence-based VET in the Netherlands: background and pitfalls. *Journal of Vocational Education and Training, 56*, 523-538.

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, MA: Kluwer Academic Publishers

Birenbaum, M., Breuer, K., Cascallar, E., Dochy, F., Dori, Y., Ridgway, J., & Wiesemes, R. (2006). 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.

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.

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

Gonczi, A. (1994). Competency based assessment in the professions in Australia. *Assessment in Education: Principles, Policy and Practice, 1*, 27-44.

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

Klatter, E. B. (2006). Competentiegerichte projectwijzers voor de lerende onderzoeker [Competence-based project books for the learning researcher]. *Develop, 2*, 24-34.

Klein, S. P., McCaffrey, D., Stecher, B., & Koretz, D. M. (1995). The reliability of mathematics portfolio scores: Lessons from the Vermont experience. *Applied Measurement in Education, 8*, 243-260.

Knight, P. T. (2000). The value of a programme-wide approach to assessment. *Assessment and Evaluation in Higher Education, 25*, 237-251.

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.

Maclellan, E. (2004). Initial knowledge states about assessment: novice teachers' conceptualisations. *Teaching and Teacher Education, 20*, 523-535.

McNamara, G., & O'Hara, J. (2005). Internal review and self-evaluation – the chosen route to school improvement in Ireland? *Studies in Educational Evaluation, 31*, 267-282.

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

Miles, M. B., & Huberman, A. M. (2003). *Qualitative data analysis. A sourcebook of new methods* (2nd ed.). Beverly Hills, CA: Sage Publications

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.

Myers, C. B., & Myers, S. M. (2007). Assessing assessment: The effects of two exam formats on course achievement and evaluation. *Innovative Higher Education, 31*, 227-236.

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

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

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 and Evaluation in Higher Education, 25*, 265-278.

Valli, L., & Rennert-Ariev, P. (2002). New standards and assessments? Curriculum transformation in teacher education. *Journal of Curriculum Studies, 34*, 201-225.

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 Education, 35*, 326-330.

Wolf, A. (1989). Can competence and knowledge mix? In J. W. Burke (Ed.), *Competency based education and training* (pp. 39-53). London: Falmer Press.

Wolf, D, Bixby, J., Glenn III, J., & Gardner, H. (1991). To use their minds well: Investigating new forms of student assessment. In G. Grant (ed.), *Review of Research in Education, Vol. 17* (pp. 31-74). Washington, DC: American Educational Research Association.

Yin, R. K. (2002). *Case study research. Design and methods* (3rd ed.), Applied Social Research Methods Series, 5. Beverly Hills, CA: Sage Publications.

7. General discussion

The aim of the research presented in this thesis was to develop, validate and explore the use of a framework of quality criteria to evaluate the quality of Competence Assessment Programmes (CAPs) in competence-based education. In the first chapter, we introduced three main research questions that were addressed in chapters 2 through 6. These questions were: (1) What quality criteria are needed to evaluate the quality of assessment programmes in competence-based education, (2) How can these quality criteria be validated, and (3) What is the utility of these quality criteria for practitioners?

This final chapter first presents the main findings and conclusions of the studies in this thesis. Then, some critical remarks and challenges for further research are presented and discussed. This chapter ends with the practical implications that can be derived from our studies.

Main findings

Research question 1: Quality criteria to evaluate CAP quality

The first main research question focused on the issue of the quality criteria that are needed to evaluate the quality of CAPs in competence-based education. This question was mainly addressed in chapter 2. First, a definition was given of Competence Assessment Programmes or CAPs, followed by a literature study on the quality criteria for assessment that are used in the testing culture (or psychometrics) and the assessment culture (or edumetrics). Chapter 2 ended with a proposal for a 10-criterion framework for CAPs. It was argued that the psychometric criteria of validity and reliability serve an important purpose in general – they serve epistemological and ethical concerns about what is being measured, and about fairness – but they should be operationalised in a different way to be suitable for CAPs and competence-based education (Benett, 1993; Linn, Baker, & Dunbar, 1991; Martin, 1997). In this thesis, reliability was worked out in the quality criteria *reproducibility of decisions* and *comparability*. Validity was not used as a separate quality criterion, but the different validity elements were incorporated in almost all quality criteria in the framework, as was shown in a systematic qualitative comparison with Messick's (1994, 1995) framework of construct validity. In addition, complementary quality criteria, derived from the assessment culture and doing

justice to the changed character of assessment in competence-based education (Dierick & Dochy, 2001; Linn et al., 1991), were suggested. Here, the comparison with Messick's framework showed that the 10-criterion framework indeed more clearly distinguished and operationalised newer competence-based quality criteria. Moreover, the stakeholders in the assessment process – teachers, students, and the work field – were explicitly included as important determinants of assessment quality, and the issue of the feasibility of carrying out the assessments was added. The literature study and comparison with Messick's framework resulted in a preliminary framework of 10 quality criteria needed to evaluate the quality of CAPs, which was validated in subsequent studies.

Research question 2: Validation of the quality criteria

The second research question pertained to the validation of this first framework of quality criteria. This was pursued in chapters 3 and 4. Chapter 3 describes how teachers working in pre-vocational and vocational education, who are the actual developers and users of many assessments, validated the framework via a questionnaire in which they gave their opinion on the importance of the different criteria for their classroom practices. The results show that teachers considered all quality criteria to be important for their own assessments, and that they consider criteria derived from psychometric and edumetric approaches to be equally important. In general, teachers in vocational education gave higher importance scores than teachers in pre-vocational education, and specifically for *costs and efficiency*, *cognitive complexity*, and *fairness*. This might be due to the fact that vocational schools have to account for the quality of their assessments to the Examination Quality Centre, a procedure which is still quite new to them and which has caused a considerable amount of stress and fear. Also, usually pre-vocational education is not the end-station of education, which means that assessment is not always seen as having a certification function. Described in chapter 4, an expert focus group meeting resulted in minor changes in the framework developed in chapter 2. Here, a group of international experts in the area of assessment and the quality of assessment validated and improved the framework in a two-day focus group meeting. The results confirmed and validated nine of the ten criteria, while three new criteria were added. Concluding, the validation process resulted in a framework of 12 quality criteria for CAPs, namely: *acceptability*, *authenticity*, *cognitive complexity*, *comparability*, *costs and efficiency*, *educational consequences*, *fairness*, *fitness for purpose*, *fitness for self-assessment*, *meaningfulness*, *reproducibility*, and *transparency*.

Research question 3: Utility of the quality criteria

The third research question into the utility of the quality criteria for practitioners was addressed in chapters 5 and 6. To this end, a self-evaluation procedure was developed, with the help of which schools evaluated the quality of their own CAP. This self-evaluation had a formative function, that is, its purpose was to stimulate reflection on CAP quality and to provide handles for improvement. It was carried

out by a group of different functionaries from a school - the department manager, an examination board member, and a teacher - who together had a full overview of the assessment programme. The 12 quality criteria were further operationalised into indicators: more concrete aspects of a quality criterion in practice, though not too detailed so they turn the self-evaluation into just ticking off a checklist.

Chapter 5 evaluated the process of this self-evaluation, and more specifically whether schools are capable of carrying out such a self-evaluation. We explored how the two phases of the self-evaluation - an individual phase carried out via a web-based tool, followed by a group interview - contributed to the self-evaluation process, and whether schools were capable of providing examples or pieces of evidence to support their opinions on CAP quality. To this end, eight schools carried out a self-evaluation of their own CAP. Results showed that the schools experienced difficulties in exactly defining their CAP (e.g., what is formative and what is summative), but that they are capable of carrying out a self-evaluation if this process is supported. The group interview appeared to be very important in the formative self-evaluation process, as different perspectives on CAP quality were aggregated there, and the participants were confronted with each others' opinions, which led to both new insights about CAP quality and spontaneous ideas for improvement. Providing support or substantiations appeared to be difficult, as most support appeared to be in the form of personal experiences, whereas very few written sources of information or empirical data were used. Regarding the utility of the quality criteria, chapter 5 therefore concludes that schools are capable of using the 12 quality criteria to evaluate the quality of their CAP, but that a clear definition of the CAP, a combination of different functionaries as evaluators, and a group interview are needed in the evaluation process.

Chapter 6 looked at the product of the self-evaluation, that is, the quality of the CAPs being evaluated. Two contrasting schools were selected for a cross-case analysis on the use of 12 quality criteria: a more 'traditional' and a more 'innovative' school. Differences between the schools showed that they seemed to operate from different frames of reference. Appearing from the examples and evidence given, differences were found for almost all 12 quality criteria, which is not surprising. The schools indeed used the quality criteria in different ways and gave different examples to account for the quality of their CAP. First, the innovative school explicitly checked whether its assessment was *transparent*, *acceptable* and *fair* in the eyes of its stakeholders, whereas the traditional school merely assumed its stakeholders to be satisfied as they expressed no complaints. Second, the innovative school explicitly designed its CAP to be *fit for purpose*, *fit for self-assessment*, and to generate positive *educational consequences*. The traditional school did not have such a clear picture in mind of the goals of competence-based education and thus could not design its CAP to stimulate these goals. Third, the schools used different approaches to assure reliability. The innovative school emphasised *reproducibility of decisions*, using two or more assessors and collecting evidence of competence in several assignments carried out in different work places. The traditional school, in contrast, emphasised *comparability* through standardisation of assessment methods, tasks and

scoring procedures. Finally, two quality criteria caused problems in both schools: *cognitive complexity* and *meaningfulness*, for which the schools could not provide any examples in their daily practice.

Concluding, with regard to the third main research question on the utility of the quality criteria for practitioners, we can say that the quality criteria, operationalised in the self-evaluation procedure, seemed to stimulate reflection on CAP quality and provide handles for improvement. Moreover, the criteria seem to enable a comparison between different schools, and offer suggestions for schools to improve their CAP. The cross-case comparison provided insight into the reasons why assessment innovations might fail or succeed.

General conclusions

Looking back at the three main research questions and the purpose of the thesis, we can conclude that the 12 quality criteria provide a valid and useful framework to systematically evaluate the quality of Competence Assessment Programmes. The idea of CAPs incorporated two of the directions in assessment research described in the general introduction: the ideas of competence-based education, and the development from testing to assessment to programmes of assessment. The other direction described in the general introduction – from psychometrics to edumetrics – was used to develop the 12 quality criteria. Incorporating these ideas, the criteria appear to do justice to the changed character of assessment in competence-based education, without neglecting the basic and important notions of valid and reliable assessments. This thesis, thus, provided new ideas about how psychometric criteria may be adapted for the use in competence-based education, and how they can be complemented with new quality criteria derived from edumetric approaches. Also, in the self-evaluation process, this thesis showed how quality criteria can be further operationalised into indicators, which adds to their utility in practice, and the transparency and understanding of what high-quality assessment programmes should look and work like in practice. This is important, because assessment quality in competence-based education is increasingly determined by the actual use of the assessments in practice, and not only by the 'correct' design of the assessment. Finally, it showed that different types of data, both quantitative and qualitative, can add to the evaluation of CAP quality, in which the value of qualitative approaches became especially clear.

Critical remarks and challenges for further research

As the general findings and conclusions show, this thesis started with the idea of evaluating the quality of assessment programmes, and a first start in this new area of research could be made. Here, we describe a number of critical remarks with regard to the studies carried out, and we discuss some challenges for further research. As a general outline for this discussion, we use the idea of the entire

Figure 7.1. Schematic representation of CAP quality evaluation process

evaluation process of CAP quality, which could be specified in five subsequent steps depicted in Figure 7.1. First, the CAP to be evaluated needs to be defined, including the methods used and their formative and summative purposes. Second, this CAP needs to be evaluated on a number of suitable and necessary quality criteria. Third, standards that specify when a CAP is 'good enough' according to these quality criteria need to be identified. Fourth, evidence needs to be collected that proves the quality of the CAP. And finally a decision needs to be made about the quality of the CAP as a whole, aggregating all evidence collected.

Some of the steps in this process have been specifically addressed in this thesis, namely step 1: CAP definition, step 2: quality criteria, and step 4: evidence collection. Here, the limitations of our studies are discussed, followed by suggestions for further research. The other steps in this process did not fall within the scope of this thesis, but are very important for further research. These are the issues of the specification of standards and the final decisions at the programme level, for which suggestions for further research are presented.

Step 1: CAP definition

This thesis focused on the quality of assessment programmes as a whole, but only a global description could be given of what assessment methods a CAP actually comprises or should comprise. We defined a CAP as a combination of both traditional and new assessment methods, that can have both formative and summative functions, and in which the actual combination of methods depends on the context and the educational programme. In addition, chapter 2 stressed that an integrated approach to competence implies that assessment methods should be used that are capable of assessing a number of elements of competence simultaneously (Gonczi, 1994; Hager, Gonczi, & Athanasou, 1994). This, in turn, implies a knowledge test alone is not sufficient, and a CAP should comprise a selection of methods such as observations, performance assessments, or a portfolio, in which different elements are integrated. Examples of such assessment programmes were given in chapter 6, which also showed that it might be dangerous to infer too much from the observation of performance, and that it may be necessary to test knowledge independently to ensure inference beyond the actual situation (Wolf, 1989).

This thesis thus provided some suggestions as to which methods should be included in a CAP, and the idea of combining traditional and new assessment methods could be substantiated by the literature. However, no prescription could be given as to what assessment methods should definitely be included in a CAP. This was mainly due to the fact that a meta-level approach was chosen, in which we asked our participants to provide evidence of the fitness for purpose of their CAP in their specific context and related to the specific content of their educational

programme. In this sense, as is described in chapter 1, we built on the national competence profiles developed for all professions, which are also used as the basis for the development of educational programmes. We did not choose a specific subject or domain to be assessed and subsequently develop standards and specific assessments for this domain, as some other authors have done, for example, to assess teacher competence (e.g., Tigelaar, Dolmans, Wolfhagen, & Van der Vleuten, 2004; Roelofs & Sanders, 2007; Van der Schaaf, Stokking, & Verloop, 2003). Further research could critically review the existing content standards for a specific profession, such as the national competence profiles for vocational education. Also, these competence profiles provide no description of how student learning should develop over time. Because new assessments take a developmental perspective, they should be guided by a developmental model of student learning (Wilson & Sloane, 2000). The specification of such a model could add to the approach taken in this thesis, and further guide assessment innovations.

Second, we did not define which methods in a CAP should be used for formative, and which for summative purposes. Some authors (Shepard, 2000; Wolf, Bixby, Glenn III, & Gardner, 1991) warn against the combination of traditional and new assessment methods in a programme, because of the danger of using only new assessment methods for formative purposes and traditional methods for summative purposes and accountability. This might cause external accountability testing to have a negative impact on classroom practices, and lead to the de-skilling and de-professionalisation of teachers, as they are not involved in high-stakes assessment issues. On the other hand, summative assessments also have a 'formative potential' (Hickey, Zuiker, Taasoobshirazi, Schafer, & Michael, 2006) in steering students' learning processes, and maybe even more so than formative assessments. Our results in chapters 5 and 6 show some problems with regard to the differentiation between formative and summative assessments, because the schools did not make such an explicit distinction, whereas the EQC only evaluates summative assessments. However, the dilemma about whether or not to combine formative and summative assessments was not resolved, and the discussion about this issue continues. At the programme level, further research is needed into the potentially negative and positive effects of combining formative and summative functions in a CAP.

Third, previous research showed that the actual assessment practices – the way they are carried out in the school - might be very different from the intended ones described on paper, or agreed on by management (Sambell & McDowell, 1998). Chapter 5 also showed that it cannot be assumed that evaluators have a shared understanding of their CAP and all assessment methods involved. Future research should therefore address how practitioners view the intended assessments. They should be involved in the development process of new assessment programmes, or at least a shared understanding should be reached on what a CAP pertains (Gulikers, Baartman, Biemans, & Mulder, in preparation). Moreover, at the programme level, research is needed into how assessment programmes are to be designed. The quality criteria presented in this thesis provide some suggestions for such guidelines, as do the discussions on methods to be included in a CAP and their

formative and summative functions. A first study on guidelines for the development of assessment programmes is currently being carried out by Dijkstra, Schuwirth, and Van der Vleuten (2007), in which guidelines are defined by experts in assessment as a first step to build a model of CAP development.

Step 2: Evaluation on quality criteria

In the second step of the evaluation process, the CAP should be evaluated on a number of suitable quality criteria. This issue was addressed in research questions 1 and 2: the development and validation of 12 quality criteria for CAPs. A first remark with regard to this framework of quality criteria pertains to the 'wheel of competence assessment' presented at the end of chapter 4. As this wheel implies, theoretical and empirical relationships exist among the different criteria. These relations, however, were not addressed or empirically tested in this thesis. As chapter 2 explains, one of the starting points of the development of the quality criteria was to keep them separated as much as possible in order to be able to provide a clear definition of each, and to prevent container concepts. Relationships among the criteria are, however, to be expected. For example, it is often assumed that more authentic and transparent assessments will improve learning (Frederiksen & Collins, 1989; Linn et al., 1991) and that having access to the criteria for judgements is necessary to assure a fair assessment (Shepard, 2000). Also, a mix of different assessment methods is assumed to increase fairness (Linn et al., 1991), and the additional time and money invested should result in more meaningful assessments that generate positive educational consequences (Hambleton, 1996). The different quality criteria could probably be grouped and put under a smaller number of headings. Also, some researchers (e.g., Arthur, Woehr, & Maldegen, 2000; Gaugler & Thornton, 1989) argue to reduce the number of criteria, as evaluators can have difficulties cognitively handling and distinguishing between many criteria. Therefore, further research is needed to explore if the number of quality criteria should indeed be reduced, and if this could be done based on different relationships between the criteria. For example, a study currently carried out (Gulikers et al., in preparation) explores whether practitioners can distinguish between all 12 quality criteria, whether the discussion of all criteria adds new information about assessment quality in a group interview, and whether a theoretical distinction between the criteria might be valuable even if practitioners do not easily make such a distinction.

Second, some critical remarks can be made regarding the validation process of the quality criteria described in chapters 3 and 4. In chapter 3, teachers' opinions were investigated through the use of a questionnaire, which yielded very high scores on all criteria. Of course, this may indicate that teachers consider the quality criteria to be important, which was corroborated by the self-evaluation studies in chapters 5 and 6. It needs to be remarked, however, that the very small differences between the quality criteria might also indicate that the questionnaire evoked socially desirable response behaviour, or that the questions were difficult to understand. In further research, therefore, teachers' opinions could be investigated using formats such as

paired comparison, in which respondents are forced to choose between two competing options, or interviews which provide a deeper insight into the reasons behind their opinions on the quality criteria. Regarding the expert focus group meeting described in chapter 4, we tried to invite experts with diverse opinions on quality issues in assessment, and to combine researchers from psychometric and edumetric backgrounds, but further research is necessary to investigate whether these findings can be generalised beyond this group. Moreover, the quality criteria could be validated with other groups of stakeholders, for example students and work field practitioners.

Finally, in this thesis, 12 quality criteria for CAPs derived from psychometric and edumetric traditions were developed and validated. Although we incorporated the basic ideas of validity and reliability in our framework, as is shown in the comparison with Messick's framework of construct validity (1994, 1995), the testing culture and the assessment culture are fundamentally different (Wolf et al., 1991), which was the reason why validity and reliability had to be operationalised in a different way. Our framework is thus more related to edumetric than to psychometric approaches. However, this does not mean that quantitative or statistical techniques such as tests of reliability or generalisability theory are not suitable to determine CAP quality. What has to be ensured is that teachers and schools can determine the quality of their assessments themselves, and can take responsibility for assessment quality (Sadler, 1987; Wolf et al., 1991). Teachers' understanding and acceptance of quality criteria for assessments is very important, as is the need for immediate feedback on the quality of their assessments (Wilson & Sloane, 2000). In this thesis, we attempted to operationalise quality criteria in such a way that teachers can work with them directly, which often meant that a more qualitative approach was taken. A different option is to provide teachers and schools with practical procedures and tools with which they can relatively easily use statistical techniques to determine assessment quality. Wilson and Sloane (2000) present an example of how to do this. They developed graphical representations of important content variables, derived from empirical analyses of student data collected in the classroom, based on multidimensional Rasch-type models. Teachers used these maps to record and track student progress and to give feedback. Research in this area could add to the theoretical approach chosen in this thesis.

Step 3: Developing standards

In this thesis we did not develop and validate standards for the 12 quality criteria, that is, 'recognised measures of what is adequate for some purpose' (Sadler, 1987, p. 194). Based on this thesis, we therefore cannot answer the question as to when a CAP is 'good enough' according to the 12 quality criteria. The development of standards can be useful because they are not only needed for summative measurements and accountability, but can also guide improvements by giving feedback about the level or standard aspired or expected (Sadler, 1987). In the literature, nearly 50 different standard-setting methods are described, all relying on human judgement (Berk, 1996). It is therefore an illusion to assume that standards possess universality in

place and time, as they are always decided upon by a group of people (Sadler, 1987). For the development of standards, Sadler advises grounding them in experience rather than theory, which implies that teachers and schools should be involved in the development of standards. A discussion of the standards and required evidence is necessary among all stakeholders involved in the assessments (Wolf et al., 1991). For the 12 quality criteria for CAPs this means that standards should be developed by schools and teachers themselves, rather than set by an external party as minimum requirements for quality. A study currently being carried out by Bronkhorst, Baartman and Stokking (in preparation) therefore investigates which standards teachers and schools implicitly use to evaluate their CAP. Here, a standard is seen as a description of an actual situation in which the participants perceive their CAP to be good enough. This study will result in a description of the standards, comparable to exemplars in performance rubrics. Further research is needed to clarify what (amount of) evidence is needed to make a 'safe' judgement on these standards, and what evidence can reasonably be collected (Hager et al., 1994; Martin, 1997).

Furthermore, research is needed into the question of whether CAPs that comply with the standards and are thus assumed to be good enough, indeed do what they intend to do and aim at. For example, some of the 12 criteria already imply other variables on which CAPs can be evaluated. For example, a good CAP should be *acceptable*, *meaningful* and *transparent*, and these criteria imply that stakeholders' opinions should be directly investigated. Also, a good CAP should stimulate the intended learning processes (*educational consequences* and *fitness for self-assessment*), which implies further research into what learning processes the CAP actually evokes. These questions also relate to the evidence that is needed to prove the quality of a CAP, which is the topic of the next paragraph. Altogether, assessments have been shown to determine students' perception of the learning environment, and students' academic achievements (e.g., Alderson & Wall, 1993; Lizzio, Wilson, & Simons, 2002; Marton & Booth, 1997). These results depend on the assessment methods used, and what they actually assess, for example deep learning or surface learning (Lizzio et al., 2002; Marton & Säljö, 1976). Future research therefore needs to address the relationships between different learning environments, students' learning conceptions, their actual learning activities and their learning outcomes, in which the assessments should comply with the 12 quality criteria and their (local) standards. A research project pursuing this goal is currently being prepared (Baartman & De Bruijn, in preparation).

Step 4: Collecting evidence

In this thesis, evidence for CAP quality was collected through a self-evaluation procedure, in which the 12 quality criteria were further operationalised into more practical indicators. However, some critical remarks need to be made with regard to the use of self-evaluations. No external measurement of CAP quality was used, other than the self-evaluations. Self-report studies have been criticised for showing only reported behaviour, instead of actual behaviour (Falchikov & Boud, 1989; Eva & Regehr, 2005). Moreover, in our self-evaluation studies we asked the participants –

department managers, examination board members, and teachers – about evidence of the opinions of other stakeholders, for example students and the work field. Our results might thus mainly show these three functionaries' perceptions of other stakeholders' opinions. No external measurement was used to directly measure students' or the work fields' opinions and verify the evidence provided in the self-evaluation. In another study (Jonsson, Baartman, & Lennung, submitted), for example, the researchers themselves collected evidence of CAP quality through documentation, student and teacher questionnaires, and analyses of assessment outcomes. Future research could, for example, ask external assessment experts to evaluate assessment quality based on written documentation (Wools, Sanders, Roelofs, & Baartman, in preparation) or compare the schools' self-evaluations to the researchers' own independent evaluation or the audits carried out by the EQC. Also, other stakeholders could be involved in the evaluation process, and their opinions could be investigated in a more direct way. For example, a study currently being carried out by Gulikers et al (in preparation) involves students and internship supervisors in the evaluation process, via a questionnaire and a group interview.

Second, the willingness to carry out a self-evaluation and collect evidence of CAP quality might be influenced by the fact that all schools participating in our self-evaluation studies were part of a national consortium working towards competence-based education. First, all schools offered educational programmes in the technical domain of laboratory technology. In further research, other domains need to be involved, as the characteristics of education and assessment might be different in other, more socially oriented domains such as social work or nursing. Second, being part of this consortium, all schools were willing to innovate their education and assessments, though their experience with and opinions of competence-based education differed, as was shown in chapter 6. Further research is needed to investigate if other schools, which might be less inclined to innovate their assessments, are willing to and capable of carrying out a self-evaluation and providing evidence of CAP quality. As actual evidence tends to confront evaluators not only with their own strengths, but also with their weaknesses, a self-critical attitude may be a necessary prerequisite for the collection and appraisal of evidence.

Third, the self-evaluations in this thesis were only used for formative purposes, which might also have been of influence on the willingness to collect positive and negative evidence. For summative purposes, some problems were identified in chapter 5, including the fact that the different schools judged their CAP from different frames of reference, making the outcomes less comparable, and the fact that the participants had difficulties defining their CAP and evaluating it on a programme level. Due to the formative nature of the evaluations in our studies, schools were willing to be self-critical, as was shown in chapters 5 and 6. When an external party such as the EQC would have been involved in the evaluation process, schools might have been less willing to participate in a study of this kind. However, if self-evaluation is to be used for summative purposes, further research is needed into the reliability and feasibility of this procedure, and into the possibly negative consequences for schools. For example, it is questionable whether self-evaluation can

be used for formative purposes (feedback on assessment quality) and summative purposes (external accountability) at the same time (e.g., Vanhoof & Van Petegem, 2007).

Finally, chapter 5 showed that when collecting evidence to substantiate or support their judgment of CAP quality, schools tend to provide personal experiences as evidence, and that in general only very few empirical data or other written sources of evidence were available. Further research should be carried out to determine what evidence is needed to make fair and feasible decisions, and what evidence practitioners and evaluators can reasonably be expected to provide. An insightful comparison on this issue was made by Gonczi (1994), who compared educational evaluation to the legal profession, in which assessors must weigh evidence to give a judgement 'on the balance of probabilities' or 'beyond reasonable doubt'. What evidence is needed depends on the context and the importance of the assessment situation – for example, formative or summative purpose – and involves professional judgement. In their study, Roelofs and Sanders (2007) conclude that it must be determined who is in the best position to supply the necessary evidence - teachers, students, parents, managers, or external experts – and that all stakeholders in the assessment should be able to provide representative and convincing pieces of evidence. Also, approaches like Kane's argument-based approach to validity, Guba and Lincoln's (1989) guidelines for constructivist evaluation, and Moss' hermeneutic approach (1994) might be helpful when looking at the evidence that is needed at the programme level. Kane (1992, 2004) describes an argument-based approach to validity, in which validity is determined by the plausibility of the arguments leading from a test score to a decision. To show the arguments are plausible, appropriate evidence needs to be collected, for example documentation or interviews with students. Guba and Lincoln (1989) describe a procedure for what they call constructivist evaluation, in which all stakeholders are explicitly involved in the evaluation process, and are assumed to have a different perspective (from a different reality) on the constructs being evaluated. Therefore, discussion on and negotiation about evidence is necessary for reaching consensus. They also describe a number of criteria for assessing the quality of this evaluation process, based on the idea of the trustworthiness of the process of gathering and appraising evidence, and the idea that evaluators should become more knowledgeable and understanding, and should be more inclined to and empowered to take actions for improvement. Moss (1994) describes similar ideas in her hermeneutic approach to assessment, in which decisions are based on textual and contextual evidence and a rational debate in which the initial interpretations are challenged and revised. Here, the transparency of the trail of evidence leading to a decision should allow external parties to evaluate the conclusions for themselves. As can be seen, most of these discussions pertain to both the collection of evidence, and the way in which a final decision is reached about quality. This final step in the evaluation process is addressed in the next paragraph.

Step 5: Reaching a final decision

Finally, when evidence has been collected, a decision needs to be made of the quality of the CAP as a whole. Evaluation at the programme level requires integrating different pieces of evidence to reach a decision. Further research should therefore address how evaluators combine these different sources of evidence into an overall judgement. Research on portfolios, in which a similar approach is used to assess students, might provide some insights here (e.g., Hamp-Lyon & Condon, 1993; Heller, Scheingold, & Myford, 1998; Van der Schaaf, Stokking, & Verloop, 2005).

Moreover, when a standard is not used as a specification of a minimum requirement for quality, we need research that explores how a fair and feasible compromise can be reached between the 12 quality criteria, which, in turn, depends on the context in which the CAP is carried out. In practice, a CAP cannot fully comply to all 12 quality criteria at the same time because potential trade-offs exist between the criteria - for example an assessment that is more authentic may become less comparable. Research needs to address how defendable decisions (i.e., a compromise) can be made depending on the specific situation, and which arguments are needed to ensure such a defendable compromise. Related to this, questions come up such as 'Should all assessments in a CAP be authentic, or is one authentic assessment enough for the programme as a whole to be authentic?' and 'If a CAP is very reproducible, should it also be very comparable?' These questions relate to compensatory and non-compensatory approaches to decision-making (Kane & Case, 2004) and compensatory and conjunctive models for scoring students' performance (Van der Schaaf et al., 2003). For CAPs, differentiation between the quality criteria might be necessary. For example, it may be reasonable to argue that all assessment methods in a CAP have to be *fit for purpose, transparent* and *fair*, and a conjunctive model might thus be most appropriate for these criteria. On the other hand, one could also argue that not all assessments in a CAP need to be *authentic*, *cognitively complex* and *fit for self-assessment*. A compensatory model might thus be most appropriate for these criteria. Moreover, these decisions also again depend on the context in which the CAP is carried out. Further research is needed to find answers to these questions, which are also related to the discussion of standards and evidence presented earlier. The steps in the evaluation process presented here are thus related to each other, and probably cannot be answered independently.

Practical implications

Finally, a number of practical implications can be derived from the studies described in this thesis, which are also related to the characteristics of assessment and the political issues in vocational education described in chapter 1.

In this thesis, 12 quality criteria for assessment programmes in competence-based education were developed, derived from psychometric and edumetric traditions. A theoretical comparison showed that these criteria do justice to the changing character of assessment in competence-based education, without ignoring

the importance of valid and reliable assessments. Comparing the 12 criteria to the standards used by the EQC, it becomes apparent that the EQC standards are more traditional, even after new standards have been developed that should be compatible with the ideas of competence-based education, and that allow more freedom to the institutions to develop their own competence-based assessments. Therefore, it may be valuable to critically review the EQC standards in light of the criteria presented and used in this thesis. Also, the self-evaluation procedure that is based on these criteria can be very helpful now that schools indeed have to carry out a self-evaluation and account for the quality of their own assessments. It can be used for different domains, and builds on the national competence profiles schools have to use to develop their educational programmes, which increases constructive alignment between learning, instruction, and assessment. Schools still have to get used to their new responsibilities, and the self-evaluation procedure developed here seems to stimulate reflection on, and improvement of CAP quality. It provides a way of systematically looking at the quality of the assessments used, and leads to professionalisation of teachers and schools.

On the other hand, this thesis also shows the difficulties of using self-evaluations for summative purposes. Vocational schools seem not yet ready and capable to do so, and a formative approach to self-evaluation – providing feedback – may be more appropriate and valuable at this stage of development. Some positive changes are already visible in this regard. After the establishment of the EQC, and after schools were given more responsibilities, the awareness of the importance of high-quality assessments seems to have increased, and many schools are taking their responsibilities and are setting up new assessment systems and quality control systems.

Another finding from this thesis pertains to the fact that the quality of assessment is determined to a great extent by its actual and correct use. It is therefore of utmost importance that teachers and other practitioners understand and can work with the assessments, and are aware of their role in assuring high-quality assessments. The implementation of new assessment methods cannot be accomplished by providing short courses for teachers. These new ideas about assessment and assessment quality should be taught in teacher education institutions. This might be the only way to assure a real shift in assessment, because the generation of students that is now enrolled in vocational education is already more used to new assessment practices, and might also be more inclined to use these practices themselves as practitioners or teachers in their future work.

Finally, the first chapter described how the EQC was discharged from its function of externally evaluating assessment quality, and how this task is now carried out by the Inspectorate of Education. Also, the new plan to develop national examinations for vocational education, suggested by the Ministry of Education in November 2007, was presented. It is not yet clear if and how the Inspectorate will change the procedures formerly used by the EQC. The plan to develop national examinations, however, does not seem to match the policies pursued in the past years. During the last five years, schools have been stimulated to develop their own

educational programmes and assessments, in collaboration with regional parties. Schools have invested a lot of time and effort in these developments, and although discussions have arisen in the media and several problems have been identified – also in this thesis – these new developments need to be given enough time to be implemented, tested, and evaluated. At this moment, research is needed on what does and does not work in competence-based education, research on what innovations lead to better learning, and research on the factors that influence successful implementation of innovations. In this respect, this thesis provided a framework of 12 quality criteria describing what constitutes high-quality assessment in competence-based education, and a practical self-evaluation procedure with the help of which schools can systematically evaluate the quality of their assessment programmes.

References

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

Arthur, W., Woehr, D., & Maldegen, R. (2000). Convergent and divergent validity of assessment center dimensions: A conceptual and empirical re-examination of the assessment center construct-related validity paradox. *Journal of Management, 26*, 813-835.

Baartman, L. K. J., & De Bruijn, E. (2007). *Does competence-based education work? The relationship between different competence-based learning environments, students' learning conceptions, and learning outcomes*. Manuscript in preparation.

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

Berk, R. A. (1996). Standards setting: The next generation (where few psychometricians have gone before!). *Applied Measurement in Education, 9*, 215-235.

Bronkhorst, L., Baartman, L. K. J., & Stokking, K. M. (2007). *Standards for the quality of competence-based assessment: Description and rationale*. Manuscript in preparation.

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.

Dijkstra, J., Schuwirth, L., & Van der Vleuten, C. M. P. (2007, November). *Programma's van toetsing* [Programmes of assessment]. Congresworkshop Nederlandse Vereniging voor Medisch Onderwijs. the Netherlands, Egmond aan Zee.

Eva, K. W., & Regehr, G. (2005). Self-assessment in the health professions: A reformulation and research agenda. *Academic Medicine, 80*, s46-s54.

Falchikov, N., & Boud, D. (1989). Student self-assessment in higher education: A meta-analysis. *Review of Educational Research, 59*, 395-430.

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

Gonczi, A. (1994). Competency based assessment in the professions in Australia. *Assessment in Education: Principles, Policy and Practice, 1*, 27-44.

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

Gulikers, J., Baartman, L. K. J., Biemans, H., & Mulder, M. (2007). Evaluating the quality of competence-based assessment by involving multiple stakeholders. Manuscript in preparation.

Hager, P., Gonczi, A., & Athanasou, J. (1994). General issues about assessment of competence. *Assessment & Evaluation in Higher Education, 19*, 3-16.

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.

Hamp-Lyon, L., & Condon, W. (1993). Questioning assumptions about portfolio-based assessment. *College Composition and Communication, 44*, 176-190.

Heller, J. I., Sheingold, K., & Myford, C. M. (1998). Reasoning about evidence in portfolios: Cognitive foundations for valid and reliable assessment. *Educational Assessment, 5*, 5-40.

Hickey, D. T., Zuiker, S. J., Taasoobshirazi, G., Schafer, N. J., & Michael, M. A. (2006). Balancing varied assessment functions to attain systemic validity: Three is the magic number. *Studies in Educational Evaluation, 32*, 180-201.

Jonsson, A., Baartman, L. K. J., & Lennung, S. (2007). *Estimating the quality of new modes of assessment. The case of an "Interactive Examination" for teacher competency*. Manuscript submitted for publication.

Gaugler, B. B., & Thornton III, G. C. (1989). Number of assessment center dimensions as a determinant of assessor accuracy. *Journal of Applied Psychology, 74*, 611-618.

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.

Kane, M. T., & Case, S. M. (2004). The reliability and validity of weighted composite scores. *Applied Measurement in Education, 12*, 221-240.

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., & Simons, R. (2002). University students' perceptions of the learning environment and academic outcomes: implications for theory and practice. *Studies in Higher Education, 27*, 27-52.

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.

Marton, F., & Booth, S. (1997). *Learning & Awareness*. Mahwah, NJ: Lawrence Erlbaum.

Marton, F., & Säljö, R. (1976). Symposium: learning processes and strategies. On qualitative differences in learning. II: Outcome as a function of the learner's conception of the task. *British Journal of Educational Psychology, 67*, 847-851.

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.

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

Roelofs, E., & Sanders, P. (2007). Towards a framework for assessing teacher competence. *European Journal of Vocational Training, 40*, 123-139.

Sadler, R. (1987). Specifying and promulgating achievement standards. *Oxford Review of Education, 13*, 191-209.

Sambell, K., & McDowell, L. (1998). The construction of the hidden curriculum: Messages and meanings in the assessment of student learning. *Assessment & Evaluation in Higher Education, 23*, 391-402.

Shepard, L. A. (2000). The role of assessment in a learning culture. *Educational Researcher, 29*, 4-14.

Tigelaar, E. H., Dolmans, D. H. J. M., Wolfhagen, H. A. P., & Van der Vleuten, C. P. M. (2004). The development and validation of a framework for teaching competencies in higher education. *Higher Education, 48*, 253-268.

Van der Schaaf, M. F., Stokking, K. M., & Verloop, N. (2003). Developing performance standards for teacher assessment by policy capturing. *Assessment & Evaluation in Higher Education, 28,* 395-410.

Van der Schaaf, M. F., Stokking, K. M., & Verloop, N. (2005). Cognitive representations in raters' assessment of teacher portfolios. *Studies in Educational Evaluation, 31*, 27-55.

Vanhoof, J., & Van Petegem, P. (2007). Matching internal and external evaluation in an era of accountability and school development: Lessons from a Flemish perspective. *Studies in Educational Evaluation, 33*, 101-119.

Wilson, M., & Sloane, K. (2000). From principles to practice: An embedded assessment system. *Applied Measurement in Education, 13,* 181-208.

Wolf, A. (1989). Can competence and knowledge mix? In J. W. Burke (Ed.), *Competency based education and training* (pp. 39-53). London: Falmer Press.

Wolf, D, Bixby, J., Glenn III, J., & Gardner, H. (1991). To use their minds well: Investigating new forms of student assessment. In G. Grant (Ed.), *Review of Research in Education, Vol. 17* (pp. 31-74). Washington, DC: American Educational Research Association.

Wools, S., Sanders, P. F, Roelofs, E. C., & Baartman, L. K. J. (2007). *Evaluatie van een instrument voor kwaliteitsbeoordeling van competentie-assessments* [Evaluation of an instrument for the quality evaluation of competence assessments]. Manuscript in preparation.

Summary

In many European countries, the ideas of competence-based education have got a firm foothold (Weigel, Mulder, & Collins, 2007), both at the level of policy making and at the level of educational practice. In the US, a similar movement towards what is called performance standards-based education can be observed (Valli & Rennert-Ariev, 2002). As a consequence of this shift towards competence-based education, assessment practices have to be changed as well, along with the ideas of what constitutes high-quality assessment. The subject of this thesis is the development, validation and practical use of a framework of quality criteria to evaluate the quality of assessment in competence-based education. Different from most research into assessment and assessment quality, this thesis focuses on assessment programmes, instead of on single assessment methods. To elucidate this idea, the concept of Competence Assessment Programmes or CAPs is introduced. In chapters 1 and 2, a CAP is described as a combination of both traditional and new assessment methods, which can have both formative and summative functions. Chapter 1 also introduces three directions in assessment research, on which the other chapters are based. First, due to the changes towards competence-based education, the content of assessments has changed and the concept of competence needs to be operationalised. Second, the assessment methods have changed from what is sometimes called a testing culture to an assessment culture, and a mix of different assessment methods in a programme is proposed. Third, the notion of what constitutes high-quality assessments has changed, in which a contrast is made between psychometric and edumetric approaches to assessment quality. The first chapter also introduces the context of this thesis: vocational education in the Netherlands, and the national consortium of vocational schools in which most studies were carried out. Finally, some political and societal issues on assessment are described, which were of influence on our studies.

The three main research questions that are addressed in chapters 2 to 6 of this thesis are:

1. What quality criteria are needed to evaluate the quality of assessment programmes in competence-based education?
2. How can these quality criteria be validated?
3. What is the utility of these quality criteria for practitioners?

The first main research question on the quality criteria that are needed to evaluate the quality of CAPs in competence-based education is mainly addressed in chapter 2, which describes a literature study on the quality criteria used in the testing culture (or psychometrics) and the assessment culture (or edumetrics). It is argued that the psychometric criteria of validity and reliability serve an important purpose in general – they serve epistemological and ethical concerns about what is being measured, and about fairness – but they should be operationalised in a different way to be suitable for competence-based education (Benett, 1993; Linn, Baker, & Dunbar, 1991; Martin, 1997). In chapter 2, reliability is worked out in the quality criteria *reproducibility* and *comparability*. Validity is not included as a separate quality criterion, but the different validity elements are incorporated in almost all quality criteria in the framework, as is shown in a comparison with Messick's (1994, 1995) framework of construct validity. Moreover, complementary quality criteria are suggested, derived from the assessment culture and doing justice to the changed character of assessment in competence-based education (Dierick & Dochy, 2001; Linn et al., 1991). A framework of 10 quality criteria is proposed, which is qualitatively compared to Messick's framework of construct validity (1994, 1995). This comparison shows that many relationships exist between psychometric and edumetric quality criteria, but that although both traditions strive for valid and reliable assessments, their operationalisation differs. The main differences seem to be that Messick (1994, 1995) mainly focuses on the technical issues of test quality whereas the new 10-criterion framework includes the stakeholders in the assessment process as important determinants of assessment quality, and that the issue of the feasibility of carrying out the assessments is added. The comparison also shows that, as expected, newer competence-based quality criteria are more clearly distinguished and operationalised in the new 10-criterion framework. Chapter 2 thus concludes that psychometric quality criteria are still valuable, but that they need to be operationalised in a different way, and complemented with quality criteria derived from the assessment culture. This results in a preliminary framework of 10 quality criteria that are needed to evaluate the quality of CAPs.

The second research question on the validation of this first framework of quality criteria is pursued in chapters 3 and 4. Chapter 3 describes how teachers working in pre-vocational and vocational education, the actual developers and users of many assessments, validate the framework through a questionnaire in which they can give their opinion on the importance of the different criteria for their classroom practices. The results show that teachers consider all ten quality criteria to be important for their own assessments. Although teachers are often thought to be reluctant towards adopting new assessment methods (Onderwijsraad, 2006; Shepard, 2000), they consider both traditional, psychometric criteria and newer, competence-based criteria to be equally important. Also, this study shows a number of differences between teachers working in pre-vocational education and teachers working in vocational education. Teachers in vocational education gave higher importance scores in general, and specifically for *costs and efficiency*, *cognitive complexity*, and *fairness*. This might be due to the fact that vocational schools have to

account for the quality of their assessments to a national Examination Quality Centre, a procedure which is still quite new to them and caused considerable stress and fear. Moreover, pre-vocational education usually is not the end-station of education, which means that assessment is not always seen as having certification as a function.

Chapter 4 describes the second validation round of the quality criteria, in which experts could give their opinion on the criteria in an expert focus group meeting. Here, a group of international experts in the area of assessment and the quality of assessment validated and improved the framework. In the focus group meeting, the experts were asked to generate all criteria they considered important for CAPs in an electronic Group Support System. This list was then discussed and compressed into a list of 13 criteria generated by the experts, which was systematically compared to the 10 criteria in the framework by means of a matrix. Results confirmed and validated nine of the ten criteria, while three new criteria were added. This validation process thus resulted in a list of 12 quality criteria for CAPs: *acceptability*, *authenticity*, *cognitive complexity*, *comparability*, *costs and efficiency*, *educational consequences*, *fairness*, *fitness for purpose*, *fitness for self-assessment*, *meaningfulness*, *reproducibility*, and *transparency*. Concluding, the findings of chapters 3 and 4 show how the framework of ten quality criteria was validated and improved, involving both research (expert focus group meeting) and practice (vocational education teachers).

The third research question into the utility of the quality criteria for practitioners is addressed in chapters 5 and 6. Chapter 5 describes the development of a self-evaluation procedure, with the help of which schools can evaluate the quality of their own CAP. This self-evaluation has a formative function, that is, its purpose is to stimulate reflection on CAP quality and to provide handles for improvement. It is carried out by a group of three functionaries from the same school – the department manager, a member of the examination board, and (another) teacher – who together have a full overview of the assessments used, both in terms of policies and regulations, and from personal practical classroom experience. The self-evaluation procedure consists of two phases: a first phase in which the participants individually evaluate their CAP by means of a web-based tool, and a group interview in which all individual evaluations are brought together and discussed. The 12 quality criteria are further operationalised into indicators: more concrete aspects of a quality criterion in practice, though not too detailed so they turn the self-evaluation into just ticking off a checklist. For each indicator, a quantitative and a qualitative judgement are given. Quantitatively, participants move an analog slide-bar from 'not at all' to 'completely'. A 'don't know' option is also available. Behind this slide-bar is a rating scale from 0 to 100, which is invisible as not to give the idea of giving a score or mark. Qualitatively, the participants support each rating by an example or a piece of evidence showing that the CAP indeed complies with the indicator. In the second phase - the group interview - the participants are asked to shortly describe their CAP, followed by a discussion on

each of the quality criteria in which they are explicitly encouraged to comment on their own and each others' ratings and examples.

Chapter 5 describes how eight vocational schools used this self-evaluation procedure to evaluate their CAP. This chapter evaluates the process of the self-evaluation, and more specifically how the individual phase and the group interview contribute to the formative self-evaluation process, and if schools are capable of providing examples or pieces of evidence to support their opinions on CAP quality. The results show that the schools experience difficulties exactly defining their CAP (e.g., what is formative, and what is summative), but that they are capable of carrying out a self-evaluation if this process is supported. The group interview appeared to be very important, as different perspectives on CAP quality are aggregated here, and the participants are confronted with each other's opinions, which leads to new insights about CAP quality and spontaneous ideas for improvement. Providing support or substantiations appeared to be difficult, as most support was in the form of personal experiences, whereas very few written sources of information or empirical data were used. Chapter 5 therefore concludes that schools are capable of using the 12 quality criteria to evaluate the quality of their CAP, but that a clear definition of the CAP, a combination of different functionaries, and a group interview are needed in this process. Moreover, self-evaluations were found to be useful for formative purposes – reflection on CAP quality, improvement of CAP quality and professionalisation – but issues of reliability may pose problems for summative purpose and accountability.

Chapter 6 looks at the product of the self-evaluation, the quality of the CAPs being evaluated. Based on the CAP characteristics of the eight schools participating in the study in chapter 5, two contrasting cases were selected for a cross-case analysis: a more 'traditional' and a more 'innovative' school. We explored (a) how these two schools use the 12 quality criteria to evaluate their CAP, whether they use different approaches to assure CAP quality appearing from different examples or pieces of evidence, (b) whether the innovative school's CAP better complies with the new, competence-based quality criteria, and (c) how differences between the schools may be explained. The results show that the two schools were willing to be self-critical towards their own assessment approaches, but emphasised that they are still in a development process towards competence-based education and assessment. They also often referred to the Examination Quality Centre, and their new responsibility to account for the quality of their examinations and to carry out a self-evaluation. Differences between the schools show that they seem to operate from different frames of reference. Appearing from the examples and evidence given, the innovative school had a much more positive attitude towards students, teachers, and innovations in general. They were also more pro-active when they encountered problems, and immediately came up with possibilities or plans for improvement. Another difference found was that the innovative school explicitly involved its stakeholders – teachers, students, and the work field – in the development and realisation of the CAP, whereas the traditional school did not know its stakeholders' opinions. These differences might be due to the fact that the innovative school has a

much more explicit vision of competence-based education, and the goals they want to achieve, while the traditional school's participants did not have a clear picture in mind of the goals of competence-based education. With regard to the 12 quality criteria, differences were found for almost all criteria, which is not surprising as two contrasting cases were chosen. The schools indeed used the quality criteria in different ways and gave different examples to account for the quality of their CAP. First, the innovative school explicitly checked whether its assessment is *transparent, acceptable* and *fair* in the eyes of its stakeholders, whereas the traditional school merely assumed its stakeholders to be satisfied as they expressed no complaints. Second, the innovative school explicitly designed its CAP to be *fit for purpose*, *fit for self-assessment*, and to generate positive *educational consequences*. The traditional school did not have such a clear picture in mind of the goals of competence-based education and thus could not design its CAP to stimulate these goals. Third, the two schools used different approaches to assure reliable assessments. The innovative school emphasised *reproducibility of decisions*, using two or more assessors and collecting evidence of competence in several assignments carried out in different work places. The traditional school, on the other hand, emphasised *comparability* through standardisation of assessment methods, tasks and scoring procedures. Finally, two quality criteria caused problems in both schools: *cognitive complexity* and *meaningfulness*, for which the schools could not provide any examples in their daily practice. Concluding, with regard to the third main research question on the utility of the quality criteria for practitioners, we can say that the quality criteria, operationalised in the self-evaluation procedure, seem to stimulate reflection on CAP quality and provide handles for improvement. Moreover, the criteria seem to enable a comparison between different schools, offer suggestions for schools to improve their CAP on the 12 quality criteria, and provide insight into the reasons why assessment innovations might fail or succeed.

Finally, chapter 7 concludes this thesis with a general discussion, in which we reflect on the studies carried out and discuss some challenges for further research. These reflections are based on the idea of the entire process of the evaluation of CAP quality, which can be divided into five steps: (1) CAP definition, (2) setting up quality criteria, (3) developing standards for these criteria, (4) collecting evidence to prove the CAP complies with the standards, and (5) reaching a decision on the quality of the CAP as a whole. Chapter 7 concludes that this thesis mainly addresses the definition of CAPs, the development of quality criteria, and the collection of evidence. Directions for further research are given for all steps in the evaluation process, and practical implications for the evaluation of CAP quality are described.

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

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.

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

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.

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.

Onderwijsraad (2006). *Doortastend onderwijstoezicht. Aanbevelingen voor toekomstig toezicht op het onderwijs. Advies uitgebracht aan het Ministerie van OC&W*. [Vigorous inspection. Recommendations for future inspection of education. Advice to the Ministry of Education]. The Hague, the Netherlands: Onderwijsraad.

Shepard, L. A. (2000). The role of assessment in a learning culture. *Educational Researcher, 29*, 4-14.

Valli, L., & Rennert-Ariev, P. (2002). New standards and assessments? Curriculum transformation in teacher education. *Journal of Curriculum Studies, 34*, 201-225.

Weigel, T., Mulder, M., & Collins, K. (2007). The concept of competence in the development of vocational education and training in selected EU member states. *Journal of Vocational Education and Training, 59*, 51-64.

Nederlandse samenvatting

Het onderwijs is in beweging. In veel Europese landen worden de ideeën van competentiegericht onderwijs ingevoerd, zowel in het onderwijsbeleid als in de dagelijkse lespraktijk (Weigel, Mulder, & Collins, 2007). In Amerika is eenzelfde verschuiving zichtbaar, hoewel soms andere termen worden gebruikt voor competenties of competentiegericht onderwijs (Valli & Rennert-Ariev, 2002). De omslag naar competentiegericht onderwijs vraagt om een verandering én in het leren, én in instructie, én in assessment. Het is dus belangrijk dat de beoordelingsmethoden of assessments aansluiten bij de ideeën over competentiegericht onderwijs. Onduidelijk is echter aan welke kwaliteitseisen assessments in competentiegericht onderwijs moeten voldoen. Bovendien zijn de opvattingen over kwaliteit van assessment veranderd met de verschuiving richting competentiegericht onderwijs.

Het doel van dit proefschrift is daarom om een aantal kwaliteitscriteria voor comptentieassessments te ontwikkelen op basis van een literatuuronderzoek, deze criteria te valideren, en het gebruik van de criteria in de praktijk te onderzoeken. Om competenties te beoordelen wordt het idee van Competentie Assessment Programma's of CAP's geïntroduceerd. Een CAP is een combinatie van zowel klassieke tests als nieuwe assessmentvormen, die zowel formatieve als summatieve functies kunnen hebben. In de hoofdstukken 1 en 2 wordt het idee van CAP's verder uitgewerkt. Hoofdstuk 1 bespreekt ook drie ontwikkelingen op het gebied van onderzoek naar assessment, die de uitgangspunten van dit proefschrift vormen. Ten eerste is dit de ontwikkeling van competentiegericht onderwijs, wat voor de inhoud van assessment betekent dat het begrip competentie verder moet worden geoperationaliseerd. Ten tweede worden nieuwe en andere assessmentvormen ontwikkeld en gebruikt. Sommige auteurs spreken hier van een overgang van een testcultuur naar een assessmentcultuur, en in dit proefschrift is daarnaast sprake van een overgang van losse assessments naar combinaties van verschillende assessmentvormen in een programma. Ten derde is het denken over kwaliteit van assessment veranderd, wat ook wel een overgang van psychometrie naar edumetrie wordt genoemd. Naast deze drie ontwikkelingen wordt in hoofdstuk 1 de context van de onderzoeken in dit proefschrift beschreven: competentiegericht beroepsonderwijs in Nederland, en de Stichting Consortium Beroepsonderwijs (www.consortiumbo.nl) waarbinnen bijna alle studies zijn uitgevoerd. Het eerste hoofdstuk eindigt met de politieke en sociale ontwikkelingen op het gebied van assessment die van grote invloed zijn op het Nederlandse beroepsonderwijs.

De drie onderzoeksvragen die aan de orde komen in dit proefschrift zijn:

1. Welke kwaliteitscriteria zijn geschikt en noodzakelijk om de kwaliteit van assessmentprogramma's in competentiegericht onderwijs te bepalen?
2. Hoe kunnen deze kwaliteitscriteria worden gevalideerd?
3. Wat de bruikbaarheid van deze kwaliteitscriteria in de praktijk?

De eerste onderzoeksvraag wordt beantwoord in hoofdstuk 2. In dit hoofdstuk wordt een literatuuronderzoek beschreven, waarmee is onderzocht welke kwaliteitscriteria voor assessment worden gebruikt in de testcultuur en de assessmentcultuur. De grondgedachte achter de ontwikkeling van de kwaliteitscriteria in dit proefschrift is om een combinatie te maken van criteria uit beide culturen. Traditionele criteria zoals betrouwbaarheid en validiteit zijn nog steeds belangrijk, maar omdat de testcultuur en de assessmentcultuur fundamenteel verschillend zijn, moeten de kwaliteitscriteria uit de testcultuur anders worden uitgewerkt voor competentiegericht onderwijs (Benett, 1993; Linn, Baker, & Dunbar, 1991; Martin, 1997). Hoofdstuk 2 beschrijft hoe het criterium betrouwbaarheid is uitgewerkt in de criteria *herhaalbaarheid van beslissingen* en *vergelijkbaarheid*. Validiteit is niet als apart kwaliteitscriterium opgenomen in dit proefschrift, maar elementen van validiteit komen terug in veel van de gebruikte kwaliteitscriteria. Dit blijkt ook uit een vergelijking van de 10 voorgestelde kwaliteitscriteria met het model van constructvaliditeit van Messick (1994, 1995). Naast deze aanpassing van de traditionele kwaliteitscriteria betrouwbaarheid en validiteit wordt ook een aantal nieuwe kwaliteitscriteria toegevoegd. Deze criteria doen recht aan de veranderde ideeën over kwaliteit van assessment in competentiegericht onderwijs (Dierick & Dochy, 2001; Linn et al., 1991), en gaan bijvoorbeeld in op betekenisvolheid en effecten op het leerproces. Op deze manier worden 10 kwaliteitscriteria voor assessments geformuleerd, die in dit hoofdstuk systematisch worden vergeleken met het model van constructvaliditeit van Messick (1994, 1995). Deze vergelijking laat zien dat in het model van Messick de nadruk ligt op technische aspecten van het ontwerp van assessments, terwijl de 10 criteria in dit proefschrift meer aandacht geven aan alle betrokkenen bij het assessmentproces, en aan de praktische uitvoering ervan. Bovendien blijkt dat, zoals verwacht, nieuwe competentiegerichte kwaliteitscriteria duidelijker naar voren komen in de 10 voorgestelde criteria. Hoofdstuk 2 eindigt daarom met de conclusie dat psychometrische criteria nog steeds belangrijk zijn – assessments moeten nog steeds betrouwbaar en valide zijn – maar dat deze criteria op een andere manier moeten worden geoperationaliseerd, en dat nieuwe kwaliteitscriteria moeten worden toegevoegd die recht doen aan de specifieke eigenschappen van assessment in competentiegericht onderwijs. In de volgende hoofdstukken wordt daarom verder gewerkt met de 10 kwaliteitscriteria zoals die vanuit de literatuur in hoofdstuk 2 zijn ontwikkeld.

De tweede onderzoeksvraag over de validering van de kwaliteitscriteria komt aan de orde in de hoofdstukken 3 en 4. In hoofdstuk 3 wordt door middel van een vragenlijst de mening van docenten in het vmbo (voorbereidend middelbaar beroepsonderwijs) en mbo (middelbaar beroepsonderwijs) gevraagd over de

kwaliteitscriteria. Uit de resultaten blijkt dat de docenten alle kwaliteitscriteria belangrijk vinden voor hun dagelijkse lespraktijk. Opvallend is dat de docenten traditionele en competentiegerichte kwaliteitscriteria even belangrijk vinden, terwijl vaak wordt gezegd dat docenten negatief staan ten opzichte van competentiegericht onderwijs (Onderwijsraad, 2006; Shepard, 2000). Ook blijkt een aantal verschillen tussen docenten werkzaam in het vmbo en het mbo. De mbo-docenten gaven hogere scores voor de belangrijkheid van alle kwaliteitscriteria, en specifiek voor *tijd en kosten*, *cognitieve complexiteit*, en *eerlijkheid*. Deze verschillen kunnen mogelijk worden verklaard door het feit dat mbo-scholen sinds 2004 verantwoording moeten afleggen over de kwaliteit van hun examens aan het KwaliteitsCentrum Examinering (KCE). Deze nieuwe verantwoordelijkheid heeft voor angst en stress gezorgd onder veel docenten. Ook wordt het vmbo vaak niet als eindonderwijs gezien, en heeft assessment daardoor minder een certificerende functie. In het algemeen konden de kwaliteitscriteria worden gevalideerd door docenten, een belangrijke groep ontwikkelaars en gebruikers van assessments.

Hoofdstuk 4 beschrijft vervolgens de validering van de kwaliteitscriteria door experts op het gebied van assessment en de kwaliteit van assessment. Deze validering vond plaats door middel van een tweedaagse focusgroepsbijeenkomst. In de bijeenkomst werd gebruik gemaakt van een elektronisch Group Support System, waarin de experts eerst zoveel mogelijk criteria invoerden die zij belangrijk vonden voor de kwaliteit van assessment. Al deze criteria werden besproken, wat resulteerde in een lijst van 13 criteria gegenereerd door de groep experts. Vervolgens presenteerden de onderzoekers de 10 kwaliteitscriteria uit het literatuuronderzoek in hoofdstuk 2. De 13 criteria van de experts en de 10 criteria uit de literatuur werden vergeleken in een matrix, waarin de experts in elke cel een cijfer gaven voor de mate van overeenkomst tussen het betreffende criterium uit de literatuur, en het door hen gegenereerde criterium. De resultaten van deze vergelijking gaven aan dat 9 van de 10 voorgestelde kwaliteitscriteria konden worden gevalideerd. Ook werden 3 criteria toegevoegd, wat resulteerde in een nieuw model van 12 kwaliteitscriteria voor CAP's: *acceptatie*, *authenticiteit*, *betekenisvolheid*, *cognitieve complexiteit*, *eerlijkheid*, *geschiktheid voor onderwijsdoelen*, *onderwijsgevolgen*, *herhaalbaarheid van beslissingen*, *ontwikkeling van zelfsturend leren*, *tijd en kosten*, *transparantie*, en *vergelijkbaarheid* Deze 12 kwaliteitscriteria vormden de basis voor de volgende hoofdstukken in dit proefschrift.

De derde onderzoeksvraag over het praktische gebruik van de kwaliteitscriteria wordt beantwoord in de hoofdstukken 5 en 6. In hoofdstuk 5 wordt de ontwikkeling van een zelfevaluatie procedure beschreven, waarmee scholen zelf de kwaliteit van hun CAP kunnen bepalen. De zelfevaluatie had een formatief doel, namelijk om reflectie op te roepen over de kwaliteit van de gebruikte assessments, en om handvatten te bieden voor verbetering van de kwaliteit. De zelfevaluatie werd uitgevoerd door drie evaluatoren met verschillende functies binnen de school: de afdelingsmanager, een lid van de examencommissie, en een andere docent. Samen hadden deze mensen een vollediger overzicht van alle gebruikte assessmentvormen dan één persoon dat zou hebben, en bovendien konden ze vanuit deze drie functies

input leveren vanuit zowel beleidsmatige vraagstukken als vanuit de dagelijkse lespraktijk. De zelfevaluatie procedure bestond uit twee fasen: in de eerste fase evalueerden de drie evaluatoren individueel hun CAP met behulp van een tool op Internet, en in de tweede fase werden alle individuele evaluaties samengenomen en besproken in een groepsinterview. Voor de zelfevaluatie procedure werden de 12 kwaliteitscriteria uit dit onderzoek verder geoperationaliseerd in indicatoren, meer concrete uitwerkingen van de kwaliteitscriteria in de praktijk. Voor elke indicator gaven de evaluatoren een kwantitatief oordeel over hun CAP door middel van een schuifje dat kon worden verschoven van "dit is helemaal niet van toepassing" naar "dit is helemaal van toepassing". Dit schuifje genereerde een score van 1 tot 100, die niet direct zichtbaar was om de evaluatoren niet het idee te geven dat ze hun CAP een cijfer moesten geven. Ook kon een knop "onbekend" worden aangeklikt. Hiernaast werd gevraagd om dit oordeel kwalitatief te onderbouwen met een voorbeeld of bewijs waaruit moest blijken dat het CAP inderdaad aan die indicator voldeed. In het groepsinterview in de tweede fase werden alle scores en bewijzen besproken en werden de deelnemers gestimuleerd te reageren op hun eigen en elkaars oordelen.

Hoofdstuk 5 gaat in op het proces van deze zelfevaluatie. Hiervoor voerden 8 scholen de zelfevaluatie procedure uit. Er werd onderzocht hoe de individuele fase en het groepsinterview bijdragen aan de zelfevaluatie, en of scholen bewijzen kunnen geven waarmee ze hun oordelen ondersteunen. De resultaten laten zien dat de scholen moeite hadden hun CAP precies te definiëren (bijvoorbeeld, wat is formatief, en wat is summatief), maar dat ze in staat zijn om een zelfevaluatie uit te voeren als ze in dit proces worden ondersteund. Het groepsinterview bleek erg belangrijk voor het zelfevaluatie proces, omdat hier de verschillende meningen en visies van de evaluatoren samenkomen, wat leidt tot nieuwe inzichten en spontane ideeën voor verbetering van het CAP. De resultaten geven ook aan dat scholen moeite hadden om bewijzen te leveren om hun standpunten mee te ondersteunen. De meeste bewijzen waren in de vorm van persoonlijke ervaringen, en er werden maar heel weinig schriftelijke bewijzen of empirische data aangedragen. De conclusie van hoofdstuk 5 is daarom dat scholen wel een zelfevaluatie kunnen uitvoeren, maar dat dit proces moet worden ondersteund, en dat een groepsinterview, en een combinatie van evaluatoren uit verschillende functies noodzakelijk lijkt. Een zelfevaluatie bleek nuttig voor formatieve doeleinden – reflectie, verbetering, en professionalisering van medewerkers – maar voor summatieve doeleinden kan de betrouwbaarheid van het proces problemen opleveren.

In hoofdstuk 6 wordt vervolgens het product van de zelfevaluaties onderzocht, namelijk de kwaliteit van de CAP's. Hiervoor werden 2 contrasterende cases geselecteerd uit de 8 scholen in hoofdstuk 5: een meer "'innovatieve" school, en een meer "traditionele" school. Deze twee scholen werden vergeleken, waarbij gekeken werd (a) hoe de 2 scholen de 12 kwaliteitscriteria gebruiken, en of ze verschillende bewijzen aandragen voor de kwaliteit van hun CAP, (b) of het CAP van de innovatieve school beter voldoet aan nieuwe kwaliteitscriteria, en (c) hoe

verschillen tussen de 2 scholen zouden kunnen worden verklaard. Uit de resultaten van deze vergelijking bleek ten eerste dat de evaluatoren van beide scholen bereid waren kritisch naar de kwaliteit van hun CAP te kijken, maar ook dat ze nog middenin het ontwikkelingsproces richting competentiegericht onderwijs zitten. Ook verwezen beide scholen regelmatig naar het KwaliteitsCentrum Examinering, en het feit dat ze verantwoording moeten afleggen voor de kwaliteit van hun examens. Dit hield beide scholen duidelijk bezig. De verschillen in bewijzen tussen de beide scholen laten zien dat de evaluatoren vanuit een verschillend referentiekader naar hun CAP keken. De innovatieve school bleek veel positiever te staan ten opzichte van leerlingen, docenten, en onderwijsvernieuwingen in het algemeen. De evaluatoren handelden ook op een meer pro-actieve wijze als ze problemen constateerden, en noemden vaak meteen een (mogelijke) oplossing voor een probleem. De traditionele school noemde ook veel problemen, maar kon hier vaak geen oplossingen aan koppelen. Verder bleek dat de innovatieve school duidelijk alle betrokkenen - leerlingen, docenten, en het werkveld - bij de assessments betrekt, terwijl de traditionele school niet expliciet naar hun mening vroeg. Deze verschillen kunnen wellicht gedeeltelijk worden verklaard door het feit dat de innovatieve school een veel duidelijker beeld heeft van competentiegericht onderwijs, en wat ze daarmee willen bereiken. De traditionele school heeft niet zo'n duidelijk beeld van de doelen van competentiegericht onderwijs. Wat betreft de 12 kwaliteitscriteria waren dan ook duidelijke verschillen zichtbaar. Ten eerste vroeg de innovatieve school expliciet aan de betrokkenen of ze het CAP *transparant*, *acceptabel* en *eerlijk* vonden, terwijl de traditionele school dit veronderstelde omdat er weinig klachten waren. Ten tweede bleek dat de evaluatoren van de innovatieve school hun CAP expliciet zo ontwierpen, dat het *geschikt was voor onderwijsdoelen*, de *ontwikkeling van zelfsturend leren* stimuleerde, en *positieve leereffecten* opriep. De evaluatoren van de traditionele school deden dit niet, juist omdat zij geen duidelijk beeld hadden van wat ze wilden bereiken met hun competentiegerichte onderwijs. Ten derde gebruikten de twee scholen verschillende manieren om betrouwbare assessments te creëren. De innovatieve school zette in op *herhaalbaarheid van beslissingen*, door meerdere assessoren te gebruiken, en meerdere bewijzen voor competentie te verzamelen. De traditionele school gebruikte veel meer het idee van *vergelijkbaarheid*, door taken, procedures en scoring te standaardiseren. Tot slot waren de ideeën van *betekenisvolheid* en *cognitieve complexiteit* bij beide scholen minder bekend. De evaluatoren konden hiervoor geen bewijzen noemen uit hun eigen CAP.

Concluderend bleek uit de hoofdstukken 5 en 6 dat de operationalisering van de kwaliteitscriteria in de zelfevaluatie procedure bruikbaar is, dat reflectie op de kwaliteit van CAP's wordt gestimuleerd, en dat handvatten worden gegeven voor de verbetering van de kwaliteit. Ook bieden de 12 criteria en hun operationalisering de mogelijkheid een vergelijking te maken tussen verschillende CAP's, wat suggesties kan bieden voor verbetering, en inzicht kan geven in de factoren voor het wel of niet slagen van innovaties.

Tot slot staat in hoofdstuk 7 van dit proefschrift een algemene discussie beschreven over de uitgevoerde onderzoeken. Hier worden alle onderzoeken

kritisch besproken, en worden suggesties gedaan voor verder onderzoek. De kritische reflectie op de uitgevoerde onderzoeken is gestructureerd aan de hand van het totale proces van de evaluatie van de kwaliteit van CAP's: (1) definitie van het CAP, (2) ontwikkeling van kwaliteitscriteria, (3) ontwikkeling van standaarden voor deze kwaliteitscriteria, (4) verzamelen van bewijzen voor de kwaliteit van het CAP, en (5) trekken van een conclusie over de kwaliteit van het CAP als geheel. Dit proefschrift besteedt vooral aandacht aan de definitie van CAP's, de ontwikkeling van kwaliteitscriteria, en het verzamelen van bewijzen hiervoor. Hoofdstuk 7 geeft suggesties voor verder onderzoek voor al deze 5 stappen van het evaluatieproces, en beschrijft de praktische implicaties van de uitgevoerde onderzoeken.

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

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.

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

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.

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.

Onderwijsraad (2006). *Doortastend onderwijstoezicht. Aanbevelingen voor toekomstig toezicht op het onderwijs. Advies uitgebracht aan het Ministerie van OC&W*. Den Haag: Onderwijsraad.

Shepard, L. A. (2000). The role of assessment in a learning culture. *Educational Researcher, 29*, 4-14.

Valli, L., & Rennert-Ariev, P. (2002). New standards and assessments? Curriculum transformation in teacher education. *Journal of Curriculum Studies, 34*, 201-225.

Weigel, T., Mulder, M., & Collins, K. (2007). The concept of competence in the development of vocational education and training in selected EU member states. *Journal of Vocational Education and Training, 59*, 51-64.

List of publications

International (refereed) journal articles

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., Bastiaens, T. J., Kirschner, P. A., & Van der Vleuten, C. P. M. (2007a). Teachers' opinions on quality criteria for Competency Assessment Programmes. *Teaching and Teacher Education, 23*, 857-867.

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

Baartman, L. K. J., Prins, F. J., Kirschner, P. A., & Van der Vleuten, C. P. M. (2007). Determining the quality of Competence Assessment Programs: A self-evaluation procedure. *Studies in Educational Evaluation, 33*, 258-281.

Submitted international (refereed) journal articles

Baartman, L. K. J., Prins, F. J., Kirschner, P. A., & Van der Vleuten, C. P. M. (2007). *Developing high-quality Competence Assessment Programmes: A cross-case analysis*. Manuscript submitted for publication.

Jonsson, A., Baartman, L. K. J., & Lennung, S. (2007). *Estimating the quality of new modes of assessment. The case of an "Interactive Examination" for Teacher Competency*. Manuscript submitted for publication.

Book chapters

Gulikers, J. T. M., Sluijsmans, D., & Baartman, L. K. J. (in press). The power of assessment in teacher education: How to assure authenticity, student involvement and assessment quality. In M. van der Klink, & A. Swennen, *Becoming a teacher educator*. Houten, the Netherlands: Springer.

Journal articles (non-refereed)

Baartman, L. K. J., Bastiaens, T. J., & Kirscher, P. A. (2005). Kwaliteitscriteria voor Competentie Assessment Programma's. *EXAMENS. Tijdschrift voor de toetspraktijk, juni*, 13-15.

Baartman, L. K. J., Prins, F. J., Kirschner, P. A., & Van der Vleuten, C. P. M. (2007). Kwaliteitsmeting van Competentie Assessment Programma's via zelfevaluatie. *OnderwijsInnovatie, maart*, 17-26.

International conference presentations

Baartman, L. K. J., Bastiaens, T. J., Kirschner, P. A., & Van der Vleuten, C. P. M. (2005, August). The Wheel of Competency Assessment. Presenting Quality Criteria for Competency Assessment Programmes. *Paper presented at the 11th biennial Conference of the European Association for Research on Learning and Instructio*n (EARLI), Nicosia, Cyprus.

Baartman, L. K. J., Bastiaens, T. J., Kirschner, P. A., & Van der Vleuten, C. P. M. (2005, August). The Wheel of Competency Assessment. Presenting Quality Criteria for Competency Assessment Programmes (nominee best PhD paper award). *Paper presented at pre-conference of the Junior Researchers of EARLI*, Nicosia, Cyprus.

Baartman, L. K. J., Prins, F. J., Kirschner, P. A., Van der Vleuten, C. P. M. (2006, August). The CAP quality meter: the evaluation and use of a self-evaluation instrument for schools to determine the quality of their Competency Assessment Programmes. *Paper presented at the 3rd biennial EARLI SIG Assessment Conference*, Northumbria, United Kingdom.

Jonsson, A., & Baartman, L. K. J. (2006, August). Estimating the quality of new modes of assessment: The case of an "Interactive Examination" for Teacher Competency. *Paper presented at the 3rd biennial EARLI SIG Assessment Conference*, Northumbria, United Kingdom.

Baartman, L. K. J., Prins, F. J., Kirschner, P. A., Van der Vleuten, C. P. M. (2006, December). Quality criteria for Competency Assessment Programmes. *Invited lecture,* University of Malmö, Sweden.

Baartman, L. K. J., Prins, F. J., Kirschner, P. A., Van der Vleuten, C. P. M. (2007, August). Determining the quality of Competence Assessment Programmes: A self-evaluation procedure. *Paper presented at the 12th biennial Conference of the European Association for Research on Learning and Instruction (EARLI)*, Budapest, Hungary.

Baartman, L. K. J., Prins, F. J., Kirschner, P. A., Van der Vleuten, C. P. M. (2007, August). Assessment in Competence-based education: Current characteristics and quality of assessments in Dutch vocational schools. *Paper presented in a symposium (organised by L. Baartman and E. Braun) at the 12th biennial Conference of the EARLI*, Budapest, Hungary.

Baartman, L. K. J., Prins, F. J., & Van Roozendaal, H. (2007, September). Quality of student assessment: 12 Quality criteria applied to an assessment programme for VET. *Paper presented at the European Conference on Educational Research (ECER)*, Ghent, Belgium.

National conference presentations

Baartman, L. K. J., Bastiaens, T. J., & Kirschner, P. A. (2004, mei). Requirements for Competency Assessment Programmes. *Paper presented at the Onderwijs Research Dagen*, Utrecht, the Netherlands.

Baartman, L. K. J., Bastiaens, T. J., Kirschner, P. A., & Van der Vleuten, C. P. M. (2005, mei). The Wheel of Competency Assessment. *Paper presented at the Onderwijs Research Dagen*, Gent, Belgium.

Baartman, L. K. J., Bastiaens, T. J., & Kirschner, P. A., Van der Vleuten, C.P.M. (2006, mei). Teachers' opinions on quality criteria for Competency Assessment Programmes. *Paper presented at the Onderwijs Research Dagen, Amsterdam*, the Netherlands.

Baartman, L. K. J., Prins, F. J., Kirschner, P. A., Van der Vleuten, C. P. M. (2007, mei). An empirical in-depth cross-case analysis of the quality of Competence Assessment Programmes. *Paper presented at the Onderwijs Research Dagen, 2007*, Groningen, the Netherlands

Other presentations

Baartman, L. K. J., Prins, F. J., Kirschner, P. A., Van der Vleuten, C. P. M. (2005, October). Kwaliteit van assessment in competentiegericht onderwijs. *Invited lecture given at the conference Beroepstaakgestuurd Leren (BGL)*, Nijkerk, the Netherlands.

Baartman, L. K. J., Prins, F. J., Kirschner, P. A., Van der Vleuten, C. P. M. (2006, March). Kwalificering. Kwaliteitscriteria voor Competentie Assessment Programma's. *Invited workshop given at the annual meeting of the Stiching Consortium Beroepsonderwijs*. Nijkerk, the Netherlands.

Baartman, L. K. J., Prins, F. J., Kirschner, P. A., Van der Vleuten, C. P. M. (2007, January). Werken met de CAP-kwaliteitsmeter. *Invited workshop given at the annual meeting of the Stichting Consortium Beroepsonderwijs*, Harderwijk, the Netherlands.

Curriculum Vitae

Liesbeth Baartman was born on the 15th of October, 1979, in Culemborg, the Netherlands. She studied Psychology / Cognitive Ergonomics at Utrecht University with a focus on educational psychology. She wrote her thesis at the TNO Human Factors Institute, concerning the design of team training scenarios. After her graduation (cum laude) in 2003, she started her PhD research at the Educational Technology Expertise Centre of the Open University of the Netherlands. She finished her PhD at Utrecht University, in close collaboration with the Open University of the Netherlands. In 2007, she received a grant (schakelbudget, 1.5 years) at Utrecht University to write a proposal and find new finances for a post doc research project. She is currently writing this proposal, teaching, and working on another research project concerning teachers' judgments of students' portfolios. Her main areas of research interest include assessment, quality issues in assessment, and learning in (competence-based) learning environments in vocational education.