Evaluation and Program Planning 34 (2011) 206–216

Contents lists available at ScienceDirect

Evaluation and Program Planning

journal homepage: www.elsevier.com/locate/evalprogplan

Self-evaluation of assessment programs: A cross-case analysis☆

Liesbeth K.J. Baartman [a,*], Frans J. Prins [a], Paul A. Kirschner [a,b], Cees P.M. van der Vleuten [c]

[a] Utrecht University, The Netherlands
[b] Open University of the Netherlands, The Netherlands
[c] Maastricht University, The Netherlands

ARTICLE INFO

Article history:
Received 14 July 2010
Received in revised form 21 January 2011
Accepted 1 March 2011

Keywords:
Assessment
Validity
Case study
Self-evaluation

ABSTRACT

The goal of this article is to contribute to the validation of a self-evaluation method, which can be used by schools to evaluate the quality of their Competence Assessment Program (CAP). The outcomes of the self-evaluations of two schools are systematically compared: a novice school with little experience in competence-based education and assessment, and an innovative school with extensive experience. The self-evaluation was based on 12 quality criteria for CAPs, including both validity and reliability, and criteria stressing the importance of the formative function of assessment, such as meaningfulness and educational consequences. In each school, teachers, management and examination board participated. Results show that the two schools use different approaches to assure assessment quality. The innovative school seems to be more aware of its own strengths and weaknesses, to have a more positive attitude towards teachers, students, and educational innovations, and to explicitly involve 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 in accordance with these goals.

High-quality assessments are an important and essential part of any curriculum. During the last two decades, many new assessments have been developed, such as performance assessments, portfolios, and workplace assessments. These assessments are supposed to measure a broader range of learning objectives (for example including practical work-based knowledge and attitudes), and generate positive effects on student learning and teaching practices (Black & Wiliam, 1998; Ecclestone & Pryor, 2003; Kirton, Hallam, Peffers, Robertson, & Stobart, 2007). In competence-based vocational education specifically, new assessments are implemented to enable the assessment of competence, the integrated use of knowledge, skills and attitudes to handle practical work problems (Brockmann, Clarke, Méhaut, & Winch, 2008). However, with the rise of these new assessments, discussions have come up about the assurance of the quality of these assessments. Traditional knowledge-focused assessments are often criticized for being too limited in scope and not being valid for assessing the nature of current learning (e.g., Linn, Baker, & Dunbar, 1991; Roth, 1998). On the other hand, the reliability of new assessments is often thought to be inadequate for high-stake purposes (e.g., Johnson, Fisher, Willeke, & McDaniel, 2003; Klein, McCaffrey, Stecher, & Koretz, 1995). With the rise of new assessments, other and complementary quality criteria for competence assessment have been suggested, such as authenticity, meaningfulness and cognitive complexity (Linn et al., 1991). These quality criteria do better justice to the character and purposes of assessment in competence-based education. For example, authenticity relates to the resemblance of the assessment to the future work situation (Gulikers, Bastiaens, & Kirschner, 2004), which fits vocational education's strive to connect workplace learning and school learning.

Besides the issue *what* quality criteria should be used, a second important issue is *how* assessment quality should be determined. This is the topic of this article. Assessment quality is usually evaluated by external authorities to satisfy accountability demands, without a real involvement of teachers and other practitioners in the school. However, the quality of new competence assessments is determined more by their actual and correct use in the classroom, and not just by their correct design (Baartman, Bastiaens, Kirschner, & van der Vleuten, 2007). This advocates the involvement of teachers and other practitioners in the assurance and improvement of assessment quality. Therefore, in a previous study we developed a self-evaluation procedure to evaluate the quality of competence assessment programs, in which teachers, examination board members and manager collaboratively evaluate the quality of their assessments (Baartman, Prins, Kirschner, & van der Vleuten, 2007). In many European countries,

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

* Corresponding author. Present address: Eindhoven School of Education, Eindhoven University of Technology, P.O. Box 513, 5600 MB Eindhoven, The Netherlands. Tel.: +31 40 2475928; fax: +31 40 2475379.
E-mail address: l.baartman@tue.nl (Liesbeth K.J. Baartman).

0149-7189/$ – see front matter © 2011 Elsevier Ltd. All rights reserved.
doi:10.1016/j.evalprogplan.2011.03.001

self-evaluation is becoming an increasingly important approach to both school improvement and accountability (McNamara & O'Hara, 2005). In the Netherlands, as in many other countries, self-evaluation has become a topic of debate since vocational schools have to demonstrate the quality of their assessments to an external quality board. One way to demonstrate this quality is to carry out a self-evaluation, but few methods exist to assist schools in carrying out a self-evaluation and schools have little experience in doing so. Therefore, the aim of this study is to validate the self-evaluation procedure developed in earlier work (Baartman, Prins, et al., 2007), by comparing the outcomes of two self-evaluations carried out by a school with extensive experience in competence-based education and assessment, and a novice school in this respect.

1. Quality criteria for Competence Assessment Programs

Because all single assessments have their limitations in what and how they assess (Chester, 2003; van der Vleuten & Schuwirth, 2005), assessment programs should be used to assess students' competences. Competence Assessment Programs (CAPs; Baartman, Bastiaens, Kirschner, & van der Vleuten, 2006) are deliberately arranged combinations of assessment methods that together form a coherent representation of what is important for competence. In these CAPs, triangulation between traditional tests and recently developed assessment methods is strived for. Traditional tests should not be ignored and discarded prematurely, because any method may contribute to the complex job of assessing competence. Therefore, van der Vleuten and Schuwirth (2005) and Baartman, Bastiaens, et al. (2007) argue that assessment should not be viewed as a psychometric problem to be solved for single assessments, but as an instructional design problem that encompasses the entire range of assessment methods used within a (part of) the curriculum.

The quality of single traditional assessments is usually evaluated using validity and reliability as quality criteria (e.g., Johnson et al., 2003), but they are not sufficient to evaluate new assessments in competence-based education. Other and complementary quality criteria have been suggested, such as the consequences, authenticity, meaningfulness and cognitive complexity of an assessment (Linn et al., 1991). Martin (1997) states that the notions of validity and reliability should change, because the notion of what constitutes an adequate assessment methods has changed with the rise of competence-based education. This does not mean, however, that assessments should not be valid and reliable anymore. Validity and reliability are not fundamentally wrong for competence assessments – they serve epistemological concerns about what is being measured, and address fairness issues – but they should be operationalized in a different way to be suitable for the often more qualitative nature and formative functions of these assessments (Benett, 1993; van der Vleuten, Schuwirth, Scheele, Driessen, & Hodges, 2010). In competence-based education, the formative function of assessment has become more important. Assessments should not only measure what learners have achieved at a certain moment in time, but should have a learning value in themselves, provide high-quality feedback and have a positive effect on future learning (Black & Wiliam, 1998; Kirton et al., 2007). They should resemble the future work situation (Gulikers et al., 2004), and focus on higher cognitive skills and thinking processes besides basic knowledge (Dierick & Dochy, 2001; Hambleton, 1996).

Also, this article argues that the quality of new assessments should be determined more by their actual and correct use in practice, and not just by their correct design (cf. Baartman, Bastiaens, et al., 2007). The quality of traditional assessments can be guaranteed by controlling development and administration. New assessments, on the contrary, can only be as good as the job done by the assessors using the instrument (van der Vleuten et al., 2010). For example, rating scales or observation criteria only work if the assessors use them correctly, understand the criteria, and conscientiously observe the student. Giving feedback requires skills that need to be trained, not just a written guidebook. Where validity and reliability could be build-in in traditional tests, for example by scrutinizing materials and standardization of procedures, assessors largely determine quality in competence assessments.

In this study, the criteria of reliability and validity were operationalized in a way as to make them more suitable for practical use. Also, they were complemented with criteria reflecting the formative goal of competence assessment, such as the meaningfulness for students' learning process, the stimulation of reflection and self-regulated learning, and the transparency of the assessment criteria and process for students (Kane, 1992, 2004; Linn et al., 1991; van der Vleuten & Schuwirth, 2005). This resulted in the 12 quality criteria for CAPs presented in Table 1, which were further operationalized into more practical indicators for the self-evaluation (see Section 2). Teachers, management and examination board were asked to provide evidence or arguments for the quality of their CAP based on these 12 quality criteria. Teachers provided arguments from their practical experience, the management provided documentation such as vision plans and instructions for teachers and

Table 1
Quality criteria for CAPs.

Criterion	Short description
Fitness for purpose	Alignment between curriculum goals and what and how is assessed. Criteria and standards should address all competences and the mix of methods should be fit to assess competence (Brown, 2004; Miller & Linn, 2000).
Self-assessment	CAPs should stimulate self-regulated learning, for example using self-assessments, and letting students formulate their own learning goals (Tillema, Kessels, & Meijers, 2000)
Comparability	Assessment tasks, criteria, working conditions and procedures should be consistent with respect to key features of interest (Baartman, Bastiaens, et al., 2007)
Reproducibility of decisions	Decisions about students should be based on multiple assessors, multiple tasks and multiple situations (Moss, 1994; van der Vleuten & Schuwirth, 2005)
Transparency	CAP should be clear and understandable for all stakeholders (Frederiksen & Collins, 1989; Linn et al., 1991)
Acceptability	All stakeholders should approve of the assessment criteria and methods (Stokking, Van der Schaaf, Jaspers, & Erkens, 2004)
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 (Dierick & Dochy, 2001; Hambleton, 1996; Linn et al., 1991)
Meaningfulness	CAPs should be learning opportunities in themselves and generate useful feedback for all stakeholders (Linn et al., 1991)
Authenticity	The degree of resemblance of a CAP to the future workplace (Gulikers, Bastiaens, & Kirschner, 2004)
Cognitive complexity	CAPs should enable the judgment of thinking process, besides assessing the product or outcome (Maclellan, 2004)
Educational consequences	The degree to which the CAP yields positive effects on learning and teaching (Messick, 1994; Schuwirth & van der Vleuten, 2004)
Costs and efficiency	The feasibility of carrying out the CAP for assessors and students (Hambleton, 1996; Linn et al., 1991)

students, and the examination board provided regulations and accountability documents. These three points of view were combined to yield a complete and overall view of assessment quality, an approach comparable to what Kane (2004) suggests in his plea for an argument-based approach to validity.

Three research questions were formulated that guided our analysis: (1) how do the innovative and novice school evaluate their CAP on the 12 quality criteria? (2) do they use the same or different approaches to assure CAP quality? and (3) does the innovative school's CAP better comply with the formative quality criteria? If the self-evaluation method, based on the 12 quality criteria described above, is valid, we should find differences in the outcomes of the self-evaluation between these two cases. Specifically, some of the 12 quality criteria focus on the formative aspects of assessment quality: acceptability, authenticity, meaningfulness, cognitive complexity, educational consequences, and self-assessment. We expected the CAP of the innovative school to better comply with these criteria, as they are explicitly connected to the ideas of competence-based education.

2. Methods

2.1. Context

As in many European countries, the ideas of competence-based education have gained a firm foothold in the Netherlands, and vocational schools are legally bound to offer a competence-based curriculum from 2010 on. Therefore, most schools are currently innovating their education and assessment practices, in which competence is seen as an integrated whole of knowledge, skills and attitudes. The reform towards competence-based education specifically aims at defining qualification profiles in more general terms linked to practical applications in the work field, instead of breaking down knowledge and skills into small 'behaviorist' units. Competence is thus defined in a different way than in for example the US.

In the Netherlands, 60 percent of all students enter vocational education, which offers 2- to 4-year courses at levels ranging from assistant worker to middle management. Three national developments are of influence on this study. First, from 2010 on, curricula and assessments have to be based on new competence-based qualification profiles, developed by representatives from social partners and vocational education. Second, the Dutch Ministry of Education, Culture and Science and the Inspectorate of Education (2007) expressed serious doubts about the quality of assessment in vocational schools and employers appeared to have little faith in the knowledge and skills of graduates entering the labor market. From 2004 on, assessment quality got special attention in the external monitoring by the Inspectorate. Third, the monitoring system itself has been adapted. Vocational schools have to carry out a self-evaluation of the quality of their assessments, which forms the basis for a more or less extensive external follow-up. Schools are given more responsibility and freedom in developing and quality assuring their assessments, provided they bear the test of the Inspectorate of Education for accountability purposes.

This study thus validates one such self-evaluation method. It needs to be noted, though, that the purpose of this self-evaluation was formative. It was meant to stimulate critical reflection on assessment quality and internal improvement and did not specifically address the requirements of the Inspectorate.

2.2. Participating schools

Two schools participated in this study. They were selected from eight schools participating in a larger research project (Baartman, Prins, et al., 2007). All offer laboratory technology education, a vocational course preparing students for a job as a laboratory assistant or laboratory technician. The schools are part of a national consortium aiming at the innovation of technical education, which started the development of problem-based lesson materials called 'unit books' in 2000 and is now developing competence-based materials called 'project books'. The competence-base materials emphasize the importance of assessment in the professional job context, the measurement of attitudes and self-regulated learning (Klatter, 2006). For this study, two contrasting cases (Yin, 2002) were selected based on their CAP characteristics. They were selected as extreme cases, that is, the most novice and most innovative school were selected, as it comes to their experience with competence-based education and assessment. It needs to be noted that the two selected schools were the extreme cases within this national consortium. The fact that they are part of such a consortium might already imply that are willing to innovate. The two extreme cases were selected based on Table 2 (see shaded portions). Five schools worked with the problem-based unit books; three worked with the competence-based project books. Furthermore, different assessment methods were used. For our comparison, school C was preferred above school B, which has similar CAP characteristics, but of which less documentation was available. School H was developing and pilot testing an entirely new CAP at the time of data collection. It needs to be noted that the two schools evaluated CAPs of different course years, which may make them less comparable in this respect. This was unavoidable because the unit books are only used in years 1 and 2, while the project books are only used in year 3. The most important differences between the two cases are described in Table 3.

2.3. Participants in the self-evaluation

In each vocational school, one teacher, one manager and one member of the examination board carried out the self-evaluation. The participating schools were asked to give the names of people who are well acquainted with the assessments used. For example, the participating teachers developed lesson and assessment materials. The functionaries did not have double roles, for example

Table 2
Summary of CAP characteristics of eight schools participating in the larger research project.

	Experience with CBE[a] (# of years)	Lesson materials	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	Portfolio
A	3	Unit books	×	×	×	×	×	×			×	
B	3	Unit books	×	×	×	×	×	×				
C	1	Unit books	×	×	×	×	×	×				
D	2	Unit books	×	×	×	×	×	×	×		×	
E	3	Unit books	×	×	×	×	×	×	×		×	
F	3	Project books		×	×	×	×	×	×	×		
G	3	Project books		×	×	×	×	×	×	×		
H	3	Project books			×	×	×		×	×	×	

[a]CBE = competence-based education.

Table 3
Summary of differences in assessment characteristics between selected cases.

Novice school's CAP	Innovative school's CAP
Teachers and managers are relative novices in the use of new assessments	Teachers and managers collaboratively designed an entirely new assessment program
Mostly knowledge tests with MC and open questions	No separate knowledge tests
Most assessments take place in the school	Most assessments take place in the workplace

as an examination board member and a teacher. Together, they have a complete overview of all assessments, in terms of national and context-specific policies and regulations, and from personal practical experience. Also, a mix of different functionaries is important to increase acceptance and ownership and different stakeholders are likely to have different perspectives on assessment quality, creating a broader and more complete picture of assessment quality (Gulikers, Baartman & Biemans, 2010; Pletinckx & Segers, 2001).

2.4. Self-evaluation method

A short description of the self-evaluation procedure is given here (for a more elaborate description, see Baartman, Prins, et al., 2007). The self-evaluation procedure consists of two phases: an individual web-questionnaire and a subsequent group interview. This was done to first stimulate the participants to apply the 12 criteria to their own CAP (following McNamara & O'Hara, 2005), and then confront the participants with each others' opinions to stimulate discussion about and reflection on assessment quality, the purpose of this (formative) self-evaluation. In the web-based questionnaire, all quality criteria were further operationalized into indicators, providing concrete quality aspects observable in practice (for a description and validation of the criteria and indicators, see Appendix A and Baartman, Prins, 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 available. Behind this slide bar was a rating scale from 0 to 100, which was invisible to avoid the idea of grading. Qualitatively, the participants supported each rating by an example from their own CAP. In the second phase, 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 2 h and had a semi-structured character. Specific questions were prepared based on the overview from the first phase. The interviewer asked for further information or explanation if the argumentation was unclear to the interviewer or the evaluators had clearly different opinions. Time-management was strictly guided to enable the discussion of all 12 criteria and all participants were encouraged to give their opinion. The group was asked to globally describe their CAP, followed by a discussion in which the participants were explicitly encouraged to comment on their own and each others' ratings and examples. Finally, besides evaluating their CAP, participants were asked to provide documentation of their assessments: policy documents describing the assessment plans and strategies, overviews of assessment methods used, scoring sheets and criteria used by the assessors, and guidelines for students and teachers.

2.5. Data analysis

Data sources were the web-based self-evaluations including quantitative scores and qualitative evidence or explanation, the transcribed group interviews and the additional documentation provided by each evaluation team. The evaluations of the two schools were systematically compared using the 12 criteria as a conceptual framework. The quantitative data were used as illustrations of the school's opinions and no statistical tests were carried out. Miles and Huberman's (2003) method of cross-case comparison was used, where qualitative data are first meaningfully reduced or reconfigured (data reduction), then organized into different displays such as diagrams or matrices (data display), from which conclusions are drawn and verified in the last phase (conclusion and verification). To answer the first two research questions, a summarizing display was constructed for each case, containing the qualitative and quantitative judgments on the 12 criteria given in the web-based questionnaire. In this display, the information of the group interviews and documentation was summarized per criterion and added in a different text color. 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 (comparable to an audit trail, see Akkerman, Admiraal, Brekelmans, & Oost, 2008). 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 were first identified by the first author, and noted down as hypotheses about general similarities and differences (e.g., the innovative school involves stakeholders, whereas the novice school does not). The first and second author then independently re-analyzed the data displays, going back to the original interviews and documentation 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 novice school appeared to do this more often). All confirmed hypotheses are presented in the results section.

3. Results

3.1. Novice school C: CAP characteristics

School C had only little experience with competence-based education (1 year). Their assessment program consists of four parts (see also Table 2, shaded portion). First, theoretical knowledge is assessed through an integral theoretical test taken at the end of each term, consisting of multiple choice questions and open questions. The school tried to organize this test around a common theme, for which all subject teachers had to develop questions, but they encountered some problems: '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'. Second, students work on 15–20 practical tasks per term (about 6–8 weeks), such as preparing a lab report or a graph with results. The products made while working on the tasks are assessed by a teacher. Three tasks per term are selected for a more thorough summative assessment. 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 while working in the laboratory at school. Students have to demonstrate all skills to a teacher, who 'ticks off' the skill on the list if it was assessed as satisfactory.

3.2. Innovative school H: CAP characteristics

The innovative school started to work with the unit books in 2003 and was developing an entirely new CAP in collaboration with regional employers at the time of data collection. The new CAP was still under construction, and although parts of it were pilot tested with students and internship supervisors, no actual user experiences were available. Here, the main part of all assessments is carried out in the workplace during internships and no separate knowledge tests are used. First, the tasks in the project books are assessed during individual internships. All students work in a company for 4 days a week, and come back to school 1 day a week to discuss the project tasks and to study the theoretical knowledge underlying the project. They work in small project teams in which the different individual tasks are combined into one large group project. Second, students' functioning in the project team is assessed using an interview. The students fill out an assessment form for themselves and their group members, which are discussed during an assessment interview with the teacher. The student sets specific learning goals for the next term. Third, the project teams present their project as a group, and questions are asked to individual team members to assess their individual contributions, focusing on the theory underlying the project. For example, students are 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 in the workplace, an overview of all competences is used, worked out in different phases of development.

3.3. General similarities and differences between the cases

In our comparison, some general similarities and differences were found that are not specific to one or two quality criteria (see Table 3). First, both schools were willing to be self-critical. They reported problems with regard to their assessments and were willing to discuss the advantages and disadvantages of their approach. They also emphasized 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. Third, many references were made to the new monitoring system used by the Inspectorate and the new responsibility of the schools to self-evaluate and account for the quality of their assessments. Both schools struggled with this new responsibility and were searching for ways to demonstrate the quality of their assessments (Table 4).

With regard to the general differences, the schools seemed 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 different. While the innovative school is quite positive, the novice school mentioned many problems, for example teachers and students having to get used to competence assessment. Second, the schools react differently in the face of problems. The innovative school mentioned possible solutions and concrete plans for improvement, while the novice school expressed uncertainty as to what and how to improve. This is also related to the third general difference, namely that the innovative school had a clear vision of competence-based education and what they wanted to achieve with their new CAP. The novice school, on the other hand, did not yet have a clear picture in mind of the goals of competence-based education, which made it difficult to make more concrete plans for improvement. Finally, the innovative school was much better informed about their stakeholders' opinions, and explicitly involved stakeholders in the assessments. The novice school implemented the unit books and associated assessment methods developed by the national consortium, without taking into account the fact that teachers were afraid their workload would increase.

3.4. Comparing the cases on the 12 quality criteria

In the next sections, the more specific similarities and differences between the innovative and the novice school for each of the 12 quality criteria are discussed. Table 5 summarizes the scores and evidence given by the participants in the self-evaluation. This table is further discussed in the sections on the different quality criteria.

3.4.1. 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. In the Dutch

Table 4
General similarities and differences between the CAP self-evaluations of the innovative school and the novice school.

General similarities	
Description	Example
Both schools are *self-critical*	'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 novice 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 improvements	'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 possible improvements mentioned]
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	'We discussed the assessments with the teachers and internship supervisors, and we piloted it with the students'	'We never explicitly measured or asked this'

Table 5
Scores and evidence of CAP quality given by the two extreme cases in the self-evaluation.

Criteria	Novice school		Innovative school	
	M	*Evidence*	*M*	*Evidence*
Fitness for purpose	59	Existing assessments are not sufficient	94	Assessments connected to new competence profile
Self-assessment	41	Important, but CAP does not stimulate	82	Explicit design of assessment to stimulate these goals. No measurement of actual effects
Educational consequences	47	No clear picture of desired learning processes	79	Explicit design of assessment to stimulate these goals. No measurement of actual effects
Comparability	96	Standardized tests to assure reliability	72	Assessment in the workplace is less comparable
Reproducibility	51	Little focus on reproducibility	80	Focus on reproducibility to assure reliability
Authenticity	71	Assessment in the school	92	Assessment in the workplace
Acceptability	71	Acceptability is assumed, but cannot be demonstrated	72	Explicit evaluation of stakeholder opinions
Fairness	84	Fairness is taken for granted	68	Personal experience with small sample
Transparency	74	Procedures and criteria specified, no check for understanding	70	Procedures and criteria specified, small checks and awareness of possible misunderstanding
Meaningfulness	50	Not enough formative assessment and feedback	71	Presumably positive, but still in development process
Cognitive complexity	56	Not specifically included in assessments	58	Needs further concretizing and attention
Costs and Efficiency	46	Time-consuming new assessments cause many problems	58	Attention was paid to feasibility in design process of new assessment program, but teachers still worry

context, both schools had to base their assessments on the new national competence-based qualification profiles. School C ($M = 59$) encountered difficulties relating its assessments to these profiles: 'It is very difficult to relate our assessments to the qualification profiles...that doesn't work anymore'. Moreover, the school was not familiar enough with the new qualification 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 H, on the other hand, co-operated with the work field to develop a new competence profile for laboratory sciences ($M = 94$). 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 struggled with the integrated assessment of knowledge, skills and attitudes. Although it referred to its integrated knowledge test, integration in this test means that knowledge questions are asked about an overarching theme, but knowledge, skills and attitudes are not actually assessed in an integrated way in a work situation. As becomes apparent from both the scores and the evidence given in the self-evaluation, 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.

3.4.2. Fitness for self-assessment

Fitness for self-assessment prescribes that CAPs should stimulate self-regulated learning, a learning goal that has become more prominent with the development of competence-based education. It is therefore not surprising that school H pays more attention to this quality criterion than school C. School H describes how their CAP stimulates self-regulated learning ($M = 82$): '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 work on in the next term'. School C did consider self-regulated learning very important, but its CAP failed to stimulate this ($M = 41$). 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 the teachers. The teacher in the self-evaluation school 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'. Apparently, 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. These differences between the cases became apparent in both the scores and the evidence in the self-evaluation.

3.4.3. Educational consequences

This criterion pertains to the effects of assessment on learning and teaching, and the curriculum as a whole. School H is much more positive than school C ($M = 79$ versus 49). It 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 does not have a clear picture of the desired learning processes in competence-based education. A few remarks made during the group interviews 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 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 organized 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 harmonized with what is necessary on the job'. Again, differences between the two cases became apparent in the self-evaluation, in both the scores and the examples given.

3.4.4. Comparability

Comparability is related to reliability as it was used for more traditional assessments. Both schools deemed comparability very important. This is interesting, as comparability is more difficult to achieve in competence-based education because less standardized assessments are used. Comparability, therefore, is worked out in different ways by both schools. The novice school ($M = 96$) 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 classroom. Explaining why they think their CAP is comparable, the school mainly referred to standardization 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'. Innovative school H could not refer to standardization, because students are assessed in different companies during their internships. It did, however, take comparability into account (M = 72): 'We do make a difference between companies ... in some, students can perform routine tasks, but not the more advanced projects in which they have to experience the entire complexity of laboratory work'. Interestingly, the school also referred 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'. This comparison shows that reliability can be achieved in different ways and that measures can be taken to assure comparability, without necessitating full standardization. While both schools give relatively high scores to their CAP, they provide different examples of how they achieved comparability.

3.4.5. Reproducibility of decisions

Reproducibility of decisions, which was already shortly referred to, is also related to reliability. Using multiple assessments and assessors, a reliable picture of a student's competences can be obtained. As described before, novice school C mainly tries to ensure reliability by standardization 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 (M = 51 versus 80). In the CAP evaluated by school C, usually 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 norms, you need only one assessor, but if you assess the student's functioning in a job situation, multiple opinions generate a more complete picture'. In contrast, reproducibility is the main way of achieving reliability for school H. It uses the competence overview to assess students during their work (multiple times), and 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 schools, comparability seems to be a more traditional way of achieving reliability, whereas reproducibility could be more suitable in competence-based education.

3.4.6. Authenticity

Authenticity relates to the resemblance of the CAP to the future job. Both schools seem to be quite satisfied with the authenticity of their CAP (M = 71 versus 92), but an interesting difference is that school C often refers to the unit books and 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 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 recognizable for the students'. School H has worked with the unit books for a number of years, now relates assessments to the actual job context: 'I think that is the strongest aspect of our new educational concept, the fact that students are actually in the workplace'. It needs to be notes that differences could be caused by the fact that school C evaluated the CAP of its first and second course 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 tasks become more complex in successive years.

3.4.7. Acceptability

Acceptability adds to the transparency criterion that stakeholders should approve of the assessments and criteria used, and have confidence in the quality of the CAP. The most salient difference between novice school C and innovative school H seems to be that school H actually involved stakeholders in the assessments, whereas school C did not ask their opinion. However, the scores between the two schools do not differ (M = 71 versus 72). School H built up its CAP from scratch and involved stakeholders from the beginning of the development process, which seemed to increase 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 the internship, and not just say, oh I think it is OK'. During the group interview of the self-evaluation, 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 get 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 out of the blue what students think of the assessments. We should ask them more specifically, interview them, or give them a questionnaire'. Interestingly, the high score of school C in the first phase of the self-evaluation seems to indicate they took the acceptability of their CAP for granted, but became aware of their inability to demonstrate this during the group interview.

3.4.8. Fairness

Fairness is related to procedures to rectify any mistakes, the use of various assessment tasks, and the fact that assessors should not be prejudiced. The results seem to show a pattern comparable to the criterion acceptability. Novice school C did not investigate whether assessors were prejudiced or not, and whether students perceive the assessments as fair: '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 exceptional'. On the other hand, school H did not 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'. This is also reflected in the relatively high scores given by both schools (M = 84 and M=68). A tentative conclusion 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 innovative school takes it for granted that its assessments are fair.

3.4.9. Transparency

Transparency prescribes that CAPs should be clear and understandable to all stakeholders, such as students, teachers and the work field. Both schools are satisfied with the transparency

of their assessments, because procedures and criteria are specified (M = 74 and 70). It is not common practice, however, to actually check whether stakeholders understand these written documents. School H carried out some checks, but also acknowledged that its CAP is completely new: '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 whether our students and supervisors really understand the assessments? I think that will take another few years'. The two schools seemed to assure transparency in different ways. Both schools reported that the assessments are discussed among teachers, but also acknowledged that their teachers are more familiar with traditional knowledge testing than with newer forms of assessment. Usually, assessment procedures and criteria are not discussed with students. The novice school assumed students to understand the assessments, because they are instructed to carefully study the guidelines and hardly ask any questions. School H referred to their pilots, 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 them in order to be able to assess themselves. 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. Like for acceptability and fairness, this criterion again shows high scores given by both schools, but a different frame of reference.

3.4.10. Meaningfulness

Meaningfulness prescribes how assessments should be meaningful learning events in themselves, for example by the feedback they generate. Both schools seem not confident that their CAP is meaningful in the eyes of students, teachers and internship supervisors, although the innovative school gives a quite high score (M = 50 and 71). School H acknowledges that an evaluation of meaningfulness is necessary some time after the new CAP has been fully implemented. School C also signals some problems. There are too few opportunities to get feedback and students do not use the opportunities they are offered, because they do not recognize 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 practice sessions for that. But it is not easy, because if you say, come to me if you have any questions, then suddenly they don't have any questions'. Also, teachers and employers do not always perceive new assessments to have an added value. They seem to be afraid that knowledge is not adequately assessed. Altogether, meaningful assessments still seem to be difficult to design and implement and more research seems warranted here. With regard to the scores given, school H seems to give its CAP a high score based on the fact that they are aware of the importance of meaningfulness, but they cannot actually demonstrate it yet.

3.4.11. Cognitive complexity

Cognitive complexity pertains to the measurement of the thinking processes professionals need to solve problems encountered on the job. Assessment should 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. Both schools referred to the lesson materials developed by the national consortium, and explained that the completion of these tasks requires thinking processes. In both courses, though, thinking processes were not explicitly assessed (M = 56 and 58). 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 C thought that 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'. These results – in terms of both scores and examples – seem to indicate that schools are still struggling with how to assess thinking processes in practice.

3.4.12. Costs and efficiency

Finally, costs and efficiency relates to the feasibility of carrying out the CAP. Whereas the novice school reports many problems, the innovative school has explicitly paid attention to feasibility in the design process of its new CAP. The scores of both schools show, however, that they are not satisfied yet (M = 46 and 58). School H involved the different stakeholders and took their opinion into account: 'We discussed how many days the internship should be ... well, to make it more cost-effective we decided for 4 days internship and 1 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'. They also noted, however, that the teachers are still worried about the feasibility of the assessments, which might explain the low score given on this criterion. School C only had a very rough idea of the time and money needed, and teachers opposed to giving feedback more regularly it would take too much time. When the school was asked if they think the investments in the CAP outweigh the positive effects, they reacted: 'At the moment there is an atmosphere of disappointment (...) 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'.

4. Conclusion and discussion

The goal of this study was to contribute to the validation of the self-evaluation method developed in earlier work (Baartman, Prins, et al., 2007), by means of comparing the outcomes of the CAP self-evaluations of a novice school and an innovative school. In case of a valid self-evaluation method, differences between the cases were expected, especially with regard to the quality criteria that focus on the formative aspects of assessment: acceptability, authenticity, meaningfulness, cognitive complexity, educational consequences, and self-assessment. All results need to be interpreted with some caution, as the self-evaluations were carried out by a small group of people, representing the other people working within their school.

With regard to CAP characteristics, the CAP of the school with more experience concerning competence-based education could indeed be regarded as 'more competence-based', as it is based upon observations in the workplace together with presentations and criterion-based interviews. It is remarkable, though, that in neither of the two schools a portfolio is used, which is generally regarded as a good instrument for the assessment of competence (e.g., Birenbaum, 1996; Dierick & Dochy, 2001). It needs to be noted, however, that the eight schools participating in the larger project were discussing the possibility of collaboratively developing a portfolio to be used by all laboratory technology schools.

Another CAP characteristic warranting discussion is the fact that the innovative school does not use any separate knowledge tests. Instead, knowledge is assessed through the work on the projects, 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). Also, Wolf, Bixby, Glenn, and Gardner (1991) state that it is dangerous to infer too much from the observation of performance and that knowledge needs to be tested independently of performance since this is the best basis for inference beyond the actual situation. Interestingly, Wolf et al. also point 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 generalizability studies (e.g., Wass, McGibbon, & van der Vleuten, 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 programs like this effectively assess students' knowledge base.

Looking at the scores and examples given for the different quality criteria in the self-evaluations of the two CAPs, differences were found for almost all criteria. First, the innovative school explicitly designed its CAP to be fit for purpose, fit for self-assessment, authentic, and to generate positive educational consequences, whereas the novice 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. On these criteria, the scores given by the innovative school are higher than the scores given by the novice school. Second, the innovative school explicitly checked whether its assessment was transparent, acceptable and fair in the eyes of its stakeholders. The novice school merely assumed their stakeholders to be satisfied as they expressed no complaints. Both schools gave their CAP high scores on these criteria, which might indicate they judge their CAP from a different frame of reference. Third, the differences between comparability and reproducibility of decisions in both scores and examples show that reliability can be assured through repeated measures by different assessors and in different contexts. Some similarities between the two cases were found as well. Although the innovative school's CAP was well thought-out, the actual effects on students' learning processes still need confirmation. 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. Here, more research seems warranted on how to develop and implement cognitive complex and meaningful assessments.

Altogether, on some quality criteria, the differences between the two cases were found as we expected them to be. The innovative school scores higher on fitness for purpose, self-assessment, educational consequences, and authenticity. The same holds for reproducibility. The fact that the innovative school scores lower on comparability can be explained by the fact that comparability and reproducibility seem to be two different ways of assuring reliability, in which reproducibility better suits a competence-based approach. For other criteria – acceptability, fairness, and transparency – no differences were found in the scores given by the two schools. Here, the scores given in the web-based questionnaire in the first phase of the self-evaluation seem to be less reliable. The group interview seems to be necessary here to get a better impression of CAP quality. For example, the novice school became aware of their inability to demonstrate acceptability during the group interview, whereas they gave their CAP a fairly high score before. A good option to solve this problem would be to ask the participants in the self-evaluation to give their CAP another score during of shortly after the group interview. This would probably give a better impression of CAP quality as all information has just been discussed during the group interview. Concluding, the results seem to support the validity of the self-evaluation to a great extent.

5. Lessons learned

This study showed the value of a self-evaluation method to evaluate the quality of Competence Assessment Programs. The results show that schools tend to score the quality of their CAP from a different frame of reference. Therefore, quantitative scoring alone seems not to be sufficient to evaluate assessment quality, neither for internal purposes, nor for accountability. For internal formative evaluation, aiming at the improvement of CAP quality, the group interview seems to be very useful, as it stimulates discussion between different stakeholder groups. For external purposes or accountability, it is not enough to give high scores. These scores need to be substantiated with evidence and examples from practice. This study showed that the novice school became aware of this fact during the group interview, while their 'overall impression' was quite positive. To stimulate schools to collect evidence of assessment quality, a formative self-evaluation procedure could start with the collaborative search for evidence, for example in student evaluations (written and oral), and systematic observations during assessments.

The combination of different evaluators as was done in this study seems to be valuable, as they perceive assessment quality in different ways, each from their own perspective as a teacher, manager or examination board member. The addition of other stakeholder groups could be even more informative, for example students, parents and representatives of the work field.

This study shows that thinking in terms of formative criteria of assessment quality – acceptability, authenticity, meaningfulness, educational consequences, self-assessment and cognitive complexity – is not common practice yet. The self-evaluation schools did deem these criteria important, but they had difficulties providing examples of the criteria and practices and sometimes did not know how to implement them. Here, self-evaluations could be valuable to increase the awareness of the importance of the formative function of assessment. Especially when it comes to accountability, this aspect of assessment is often overlooked.

Appendix A. Twelve quality criteria and their indicators[a]

Acceptability		**Costs and efficiency**		**Fitness for self-assessment**	
1	Students approve of criteria	1	Time and money estimated	1	Self- and peer-assessment
2	Students approve of procedure	2	Deliberately choosing mix	2	Giving and receiving feedback
3	Teachers approve of CAP	3	Yearly evaluation of efficiency	3	Reflection on personal development
4	Employers approve of CAP	4	Positive effects outweigh investments	4	Formulation of personal learning goals
5	Confidence in quality CAP				
Authenticity		**Educational consequences**		**Meaningfulness**	
1	Assessment tasks resemble job	1	Desired learning processes stimulated	1	Feedback formative useful
2	Working conditions resemble job	2	Positive influence on students	2	Feedback summative useful
3	Social context resembles job	3	Positive influence on teachers	3	Assessment is opportunity to learn
4	Assessment criteria resemble job	4	Improved if negative effects	4	Students think criteria meaningful
		5	Curriculum adapted if CAP warrants	5	Teachers/employers think criteria meaningful
Cognitive complexity		**Fairness**		**Reproducibility of decisions**	
1	Tasks trigger thinking steps	1	Procedures to rectify mistakes	1	Several times
2	Explain choices	2	Weights based on importance	2	Several assessors
3	Criteria address thinking steps	3	Assessors not prejudiced	3	Assessors with different backgrounds
4	Tasks require thinking level	4	Various types of assessment tasks	4	Equal discussion between assessors
		5	Student think CAP is fair	5	Trained and competent assessors
				6	Several work situation
Comparability		**Fitness for purpose**		**Transparency**	
1	Assessment tasks comparable	1	Coverage of competency profile	1	Student know formative of summative
2	Working conditions comparable	2	Integrated assessment of K/S/A	2	Students know criteria
3	Assessment criteria comparable	3	Mix of different assessment forms	3	Students know procedures
4	Assessment procedure comparable	4	Both summative and formative forms	4	Teachers know and understand
		5	Forms match with educational goals	5	Employers know and understand
				6	External party can audit

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

References

Akkerman, S., Admiraal, W., Brekelmans, M., & Oost, H. (2008). Auditing quality of research in social sciences. *Quality & Quantity, 42*, 257–274.

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., 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.

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.

Birenbaum, M. (1996). Assessment 2000: Towards a pluralistic approach to assessment. In M. Birenbaum & *Alternatives in assessment of achievement, learning processes and prior knowledge* (pp. 3–29). Boston, MA: Kluwer Academic Publishers.

Black, P., & Wiliam, D. (1998). Assessment and classroom learning. *Assessment in Education, 5*(1), 7–74.

Brockmann, M., Clarke, L., Méhaut, P., & Winch, C. (2008). Competence-based vocational education and training (VET): the cases of England and France in a European perspective. *Vocations and Learning, 1*, 227–244.

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.

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.

Ecclestone, K., & Pryor, J. (2003). 'Learning careers' or 'assessment careers'? The impact of assessment systems on learning. *British Educational Research Journal, 29*, 471–488.

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

Gulikers, J. T. M., Baartman, L. K. J., & Biemans, H. J. A. (2010). Facilitating evaluations of innovative, competence-based assessments: Creating understanding and involving multiple stakeholders. *Evaluation and Program Planning, 33*, 120–127.

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.

Johnson, R. L., Fisher, S., Willeke, M. J., & McDaniel, F., II (2003). Portfolio assessment in a collaborative program evaluation: The reliability and validity of a family literacy portfolio. *Evaluation and Program Planning, 26*, 367–377.

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.

Kirton, A., Hallam, S., Peffers, J., Robertson, P., & Stobart, G. (2007). Revolution, evolution or a Trojan horse? Piloting assessment for learning in some Scottish primary schools. *British Educational Research Journal, 33*(4), 605–627.

Klatter, E. B. (2006). Competentiegerichte projectwijzers voor de lerende onderzoeker [Competence-based project books for the learning researcher]. *Develop, Kwartaaltijdschrift voor Human Resources Development, 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.

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.

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.

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.

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

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

Ministry of Education, Culture and Science. (2007, September). *Examens MBO en positie KCE (Exams in VET and the position of the Examination Quality Centre)* Retrieved January 20, 2009, from http://www.minocw.nl/documenten/16638d.pdf.

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

Pletinckx, J., & Segers, M. (2001). Programme evaluation as an instrument for quality-assurance in a student-oriented educational system. *Studies in Educational Evaluation, 27*, 355–372.

Roth, W. M. (1998). Situated cognition and assessment of competence in science. *Evaluation and Program Planning, 21*, 155–169.

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.

van der Vleuten, C. P. M., Schuwirth, L. W., Scheele, F., Driessen, E. W., & Hodges, B. (2010). The assessment of professional competence: Building blocks for theory development. *Best Practice & Research Clinical Obstetrics and Gynaecology* doi:10.1016/j.bpobgyn.2010.04.001.

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

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

Liesbeth K.J. Baartman is a postdoctoral researcher at Eindhoven University of Technology, Eindhoven School of Education. She wrote her PhD on quality criteria for assessment programs in competence-based education. Currently, her research focuses on competences and competence assessment in science and technology education.

Frans J. Prins is an assistant professor at Utrecht University, department of Pedagogical and Educational Sciences. His areas of expertise are assessment, metacognition, inductive learning, peer feedback and learning styles.

Paul A. Kirschner is a professor of education, program director of the learning and cognition programme of the Centre for Learning Sciences and Technology of the Open University of the Netherlands. His areas of expertise include lifelong learning for professional and personal development, information and communication technology in education and assessment of skills and competence.

Cees P.M. van der Vleuten is a professor of education, Chair of the Department of Educational Development and Research and Scientific Director of the School of Health Professions Education, Faculty of Health, Medicine and Life Sciences of Maastricht University, the Netherlands. His research focuses on a programmatic approach to assessment.