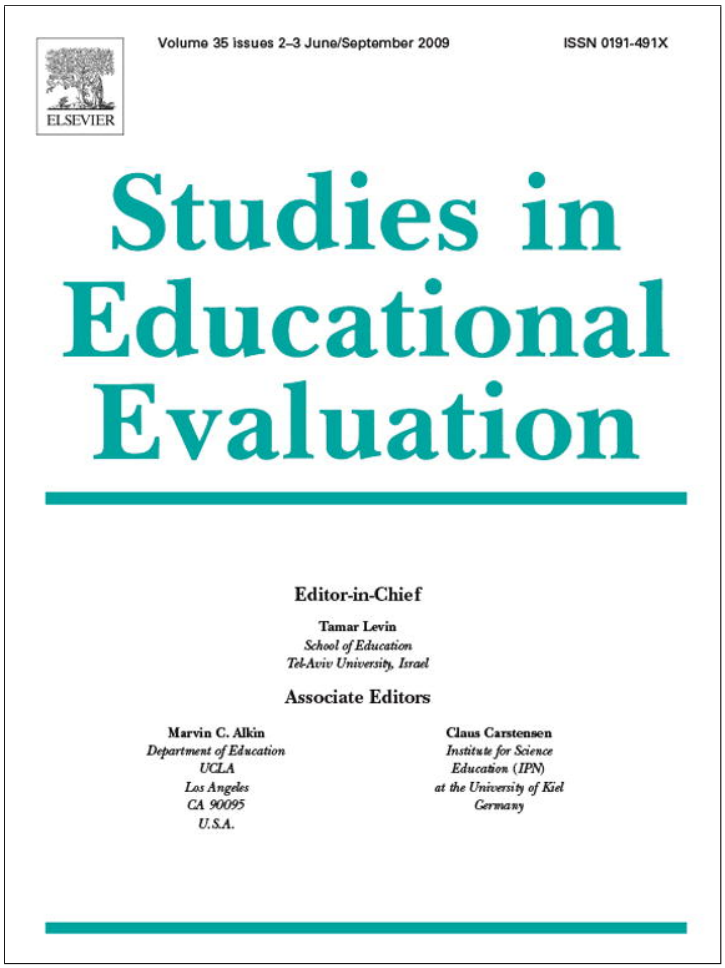

Studies in Educational Evaluation 35 (2009) 110–119

Contents lists available at ScienceDirect

Studies in Educational Evaluation

journal homepage: www.elsevier.com/stueduc

Developer, teacher, student and employer evaluations of competence-based assessment quality

J. Gulikers *, H. Biemans, M. Mulder

Education and Competence Studies, Wageningen University, P.O. Box 8130, 6700 EW Wageningen, The Netherlands

ARTICLE INFO

Keywords:
Assessment quality
Competence-based assessment
Student evaluation
Perceptions
Stakeholders
Vocational education and training

ABSTRACT

This study examines how different stakeholders experience the quality of a nationally developed assessment framework for summative, competence-based assessment (CBA) in AVET, which aims to reflect theoretical characteristics of high quality CBAs. The quality of two summative CBAs, based on this national framework, is evaluated along an extensive, validated set of quality criteria for CBA evaluation and through involving key stakeholders (i.e., students, teachers, developers, and employers). By triangulating quantitative and qualitative evaluations and argumentations of key stakeholders, this study gives insight into the processes and characteristics that determine CBA quality in VET educational practice in relation to theoretical notions of high quality CBAs. Results support many theoretical characteristics and refine them for reaching quality in actual assessment practice. Strikingly, developers and teachers are more critical about the assessment quality than students and employers. The discussion reflects on the theoretical CBA characteristics in the light of the empirical findings and deduces practical implications for the national assessment framework as well as other summative CBAs in VET.

Introduction

Professional education aims at preparing students for effective functioning in the profession. Educational assessments should therefore correspond to what is expected from students in the world of work (Gulikers, Bastiaens, & Kirschner, 2004; Kaslow et al., 2007). Indeed, various new modes of assessment practices are developed to comply with professional requirements, often called 'competence-based assessments' (CBAs). These are performance-based instead of purely knowledge-based measurements, requiring students to perform professional tasks; they place emphasis on generic transferable competencies relevant across professions instead of only focusing on discipline specific knowledge (Gulikers, Bastiaens, Kirschner, & Kester, 2006; Kaslow et al., 2007). CBAs are also more often conducted in the workplace (Smith, 2007; Strickland, Simons, Harris, Robertson, & Harford, 2001) and also pay attention to students' ability to critically reflect upon their future professional practice and performance.

Problematic, however, is that these assessments frequently are developed based on common sense or intuition instead of scientific or empirical evidence about effective, high quality CBAs (e.g., Cummings & Maxwell, 1999). Baker (2007) stressed the importance of critically examining the quality of our new assessments and looking beyond the traditional school boundaries to create greater connections between school and the workforce to build assessments of higher quality.

Characteristics of high quality competence-based assessments

It becomes widely recognized that new assessments or CBAs have different characteristics than traditional, standardized, written tests aiming at testing a knowledge base (e.g., Segers, Dochy, & Cascallar, 2003). Many theoretical notions have been put forward to characterize CBAs, like focusing on performance in various authentic situations, combining multiple methods, involving multiple assessors preferably with different backgrounds, using criterion-references scoring, and integrating learning with assessment activities. A preview to the first two columns of Table 1 shows an overview of the characteristics mentioned by many researchers and the reasons for their importance (e.g., Baartman, Bastiaens, Kirschner, & van der Vleuten, 2006; Birenbaum et al., 2006; Dierick & Dochy, 2001; Grainger, Purnell, & Zipf, 2008; Gulikers et al., 2004; Harlen, 2005; Johnston, 2004; Kaslow et al., 2007; Leigh et al., 2007; Schuwirth & van der Vleuten, 2006; Segers et al., 2003).

Unfortunately, there is still little empirical evidence on the quality of CBAs that incorporate these theoretical characteristics (Segers & Dochy, 2006). Existing research focuses on examining specific characteristics like authenticity (Gulikers et al., 2006) or student involvement (Sluijsmans & Prins, 2006) instead of

* Corresponding author. Tel.: +31 317 484332; fax: +31 317 484573.
E-mail address: judith.gulikers@wur.nl (J. Gulikers).

0191-491X/$ – see front matter © 2009 Elsevier Ltd. All rights reserved.
doi:10.1016/j.stueduc.2009.05.002

examining a CBA in its whole width and coherence, or focus on the effects of a certain assessment on students' study approaches (e.g. Harlen, 2005). There is still little empirical evidence showing what theoretical characteristics of CBA actually impact the quality of CBAs in practice. This issue is even further complicated by the acknowledgement that these new assessments require a new way of examining their quality.

Examining assessment quality: other quality criteria and other processes

As CBAs differ from traditional knowledge tests on many fundamental aspects, they necessitate a new way of examining their quality (Baartman et al., 2006; Benett, 1993; Birenbaum, 2007; Dierick & Dochy, 2001; Linn, Baker, & Dunbar, 1991; Messick, 1994). This holds for both the evaluation criteria as well as the process of examining quality. Psychometric quality criteria like reliability and validity remain important in new assessments, but their operationalisation should change in line with the new notions of competence-based assessments (Benett, 1993). Moreover, researchers have proposed additional new quality criteria that should be incorporated in a quality framework in order to address specific new characteristics of competence-based assessment (see also Table 1) that are not addressed in the psychometric framework. These are criteria like authenticity, transparency, or educational consequences (e.g., Linn et al., 1991; Messick, 1994).

Also the process of examining assessment quality and the kind of evidence required as arguments for assessment quality are changing (Birenbaum, 2007; Kane, 2008). Researchers argue that assessment quality is not purely an inherent aspect of the assessment method, it largely depends on how this method is actually implemented in a certain educational context (Kane, 2008); whether or not it is perceived to have good quality by all involved stakeholders, including students and employers (Birenbaum, 2007; Gulikers et al., 2004; Struyven, Dochy, & Janssens, 2003), and how this assessment, and students' perception thereof, affects students learning and motivation (e.g., Messick, 1994). As a result, a growing number of researchers make a plea for: (a) more qualitative argumentation for assessment quality based on how an assessment method is actually used in educational practice, instead of only examining the quality of the assessment instrument as such, and (b) for the involvement of multiple stakeholders and their experiences in the evaluation process. Different stakeholders might have different perspectives on the quality of a certain CBA and combining these perspectives results in a more valid and complete picture of the actual quality of the assessment (Birenbaum, 2007; Kane, 2008; Guba & Lincoln, 1989). Both agreement as well as differences between stakeholders' perceptions of assessment elements signal important quality issues of the CBA.

Research questions

Increasingly, research and policy agendas stress the need for evidence-based research on what works and what does not in innovative educational practices like competence-based education (Slavin, 2008; Van der Vleuten & Schuwirth, 2005). Therefore, the research questions in this study are: (1) How do different stakeholders (developers, teachers/assessors, students, and employers) experience the quality of a CBA that is developed along the theoretical characteristics of high quality assessments? And (2) what arguments do stakeholders provide for justifying their quality evaluations. By answering these research questions, this study aims to find empirical evidence for the theoretical characteristics of CBA and their relationship to CBA quality criteria.

Context of the study: vocational education and training

These research questions will be answered through examining the quality of two CBAs in senior secondary Agricultural Vocational Education and Training (AVET) in the Netherlands. Both CBAs are based on the same national assessment framework developed in AVET that aims to reflect many theoretical characteristics of high quality CBAs (see Table 1). Vocational Education and Training (VET) in the Netherlands, educating 42% of the student population, is a practically and occupationally oriented type of education in which learning and working are intertwined. To meet labor market objections, VET schools are obliged by the government to have competence-based curricula and assessments by 2010. A standard set of 25 generic competencies for VET has been developed, based on the universal SHL competency framework (www.shl.com) (e.g., 'collaborating and consulting', 'applying professional knowledge', or 'planning and organizing'). Based on this framework, national qualification profiles have been developed for all educational VET trajectories concretizing these broad SHL competencies into a number of core job tasks for a certain VET trajectory (e.g., preparing and organizing meetings for a secretary or providing care for patients for a nurse assistant). Schools are given the responsibility to develop CBAs to assess students along the qualification profile. Summative CBAs aim to assess and accredit *all* learning in VET in an integrated way. This is different from using practical or apprenticeship assessments, that only assessed placement learning next to separate knowledge and skills test for in-school learning (Smith, 2007; Strickland et al., 2001). Obviously, the quality of these all including summative CBAs and their recognition by students and employers is a pressing issue in this context.

In this study the quality of two of these all including summative CBAs in AVET are evaluated along an extensive and validated set of quality criteria for CBA evaluation (Baartman, Bastiaens, Kirschner, & van der Vleuten, 2007a; Baartman, Bastiaens, Kirschner, & Vleuten, 2007b) and through involving key stakeholders. By triangulating quantitative and qualitative evaluations and argumentation of key stakeholders on all quality criteria, this study aims to gain insight into the processes and characteristics that determine CBA quality in VET educational practice in relation to theoretical notions of high quality CBAs and the national assessment framework for AVET.

Method

Context: the national assessment framework

AVET institutions are precursors in the Netherlands with respect to competence-based curricula and they are developing CBAs through national collaborative initiatives. Teachers from all AVET institutions ($n = 12$) and representatives of the workfields collaboratively developed a national assessment framework, based on the theoretical characteristics of high quality CBAs (Table 1), for assessing all agricultural competency profiles (e.g. gardener, florist or animal care specialist). This framework is recognized as a quality assessment by the accrediting body at the national level. In short, this framework described three basic elements for every CBA:

- Content: a critical job situation (CJS) for a specific AVET competency profile describing a professional situation that includes several professional tasks and dilemmas. A number of specific and generic competencies needed to successfully perform this CJS are also described;
- Methods: the CBA should consist of two elements, being a performance assessment-on-the-job observed by two assessors (i.e., one teachers and one employer) and a criterion-based interview (CBI). With his/her performance-on-the-job and given

Table 1
Theoretical CBA characteristics and their operationalisation in the National Assessment Framework.

	Theoretical characteristics	Theoretical explanation or reasons	Operationalisation in national assessment framework of AVET
1.	Contextualized in professional practice	Resembling real professional practice in activities, context and thinking processes and assessment criteria Assessing true professional competence requires measurements of performing professional tasks in the real, complex professional world (e.g., Benett, 1993; Gulikers et al., 2004; Segers et al., 2003)	Critical job situation (CJS) is starting point of the assessment Holistic overall assessment criterion related to job performance = 'can the student perform the CJS in real life?' Assessment conducted in professional practice (work placement context of every individual student) Involves actual performance of professional tasks and dealing with upcoming professional dilemmas
2.	Collaboration with/involvement of work-field	Developing and conducting the assessment should involve practitioners (e.g., Baker, 2007; Gulikers et al., 2007)	Involved in development of competency profile and assessment. National assessment framework and content validated by workfield Employers involved as co-assessors
3.	Incorporation of multiple methods/moments that address product and process	Assessing the complexity of competencies requires a combination of assessment methods addressing competence in different situations. Competence implies flexibility: more attention to the process of solving a problem next to the actual solution (=product) (Baartman et al., 2006; Kaslow et al., 2007; Linn et al., 1991).	Combination of two methods: Observation of performance-on-the-job (= product and process), Criterion-based interview: motivating performance (= process) Both methods are conducted at a fixed time period after learning
4.	Multiple assessors, preferably with different backgrounds	Assessors with different backgrounds have different reference frames for judging the same performance. 'The truth is a matter of consensus' (Johnston, 2004). Inter-subjectivity, instead of objectivity (Baartman et al., 2006; Benett, 1993; Schuwirth & van der Vleuten, 2006)	At least two assessors: one teachers, one employer
5.	Addressing higher-order processes, including reflection and/or self assessment, and the ability to transfer to new situations	Competent performance in complex world requires many higher-order thinking processes and flexibly using them in various situations Professional performance requires performing professional tasks, but also reflection in and on action (Schön, 1987) Stimulating life-long learning skills by incorporating self-assessment (Baartman et al., 2006; Birenbaum et al., 2006; Dierick & Dochy, 2001)	Explicitly stated goals of the Criterion-based interview: motivating choices made in performance-on-the-job, reflecting on action in performance-on-the-job, and addressing transfer to new situations Self-assessment is not mentioned in this assessment framework
6.	Integrated with instruction	To stimulate required learning processes, instruction and assessment should address the same competencies and learning processes (Birenbaum et al., 2006; Dochy, 2005; Gulikers et al., 2004).	Schools are free in the way they set up their curriculum. There is no obligatory or explicitly described curriculum preceding the assessment
7.	Individualization of assessments	Assessment should allow for differentiation between students to be responsive to students' needs and situations (Dierick & Dochy, 2001; Segers et al., 2003)	Every student conducts the assessment in his/her own work placement context, Different students are assessed by different employer-assessor
8.	Increased student responsibility and involvement	Students should be given more responsibility over the content, form, and timing of their assessment Students should be involved as co-developers and/or co-assessors (Biemans et al., 2004; Gulikers et al., 2004; Sluijsmans & Prins, 2006)	Students are not given explicit responsibilities, the assessment is guided by the assessors. Students are not involved as developers and/or assessors
9.	Combining assessment *of* and assessment *for* learning	Feedback is crucial for making assessment a learning experience (Birenbaum et al., 2006; Harlen, 2005) Also summative assessments should inform further learning, development and teaching (formative purpose)	Strict separation between summative and formative functions: the CBA is not developed to have a formative purpose Feedback is not incorporated as part of the assessment
10.	Criterion-referenced scoring	Evaluating against a required level of competence (=criteria/standards) instead of comparing students (norm-referenced) Literature shows debate about appropriate level of detail of criteria (Grainger et al., 2008; Johnston, 2004)	Criterion-referenced: overall holistic and dichotomic criterion "is the student able to competently perform the CJS in real professional practice: Yes or no" Explicit instruction not to tick of individual competencies or activities

Table 1 (*Continued*)

	Theoretical characteristics	Theoretical explanation or reasons	Operationalisation in national assessment framework of AVET
11.	Transparency of assessment	Assessment and its criteria should be known beforehand for all participating parties, including students, as this guides student learning (Dierick & Dochy, 2001; Gulikers et al., 2004, 2008)	The national framework, filled in for the specific competency profile, including all competencies (with expected performance levels) and assessment procedures were provided to all parties from the start

argumentations in the CBI the student has to prove to, at least, two assessors that he/she is competent in performing the CJS in its whole width and coherence. A combination of two, three or four[1] of these CBAs together cover all critical job situation of an AVET qualification profile and constitute the summative assessment of this AVET trajectory;

- Purpose: summative, and *not* formative. Based on students' performance and CBI, the two assessors have to holistically judge the students' competence on one crucial criterion being: 'Is this student competent in performing the CJS in real professional practice or not?' This holistic judgment depends on the professional expertise of the assessor(s), instead of on ticking of a list of more detailed assessment criteria.

The national assessment framework in relation to theoretical CBA characteristics

The right column of Table 1 displays more in-depth how the theoretical CBA characteristics were filled in this national assessment framework. Many theoretical characteristics were given high priority in the national assessment framework: much emphasis on a strong resemblance between the assessments and the professional field, strong collaboration with the work field, combination of two assessment methods, use of at least two assessors from different backgrounds, attention for reflection and transfer, and high transparency for all groups. Three characteristics were not followed up: the national assessment did not emphasize increased student responsibility, it stresses a strict separation between summative and formative assessment (goals), and does not provide any information or requirements for the integration of the assessment with the curriculum. These characteristics were expected to be either not suitable for the VET context (student responsibility) or were outside of the scope of the national assessment framework. The remaining two characteristics (individualization and criterion-referenced assessment) were partly incorporated.

The actual assessments

The nationally described CBA is still a written product. Every assessment development team within an AVET school has to work out this written product into an actual assessment based on their own context, wishes, requirements and possibilities. The national framework sets out several obligatory elements (content in the CJS and minimal procedural guidelines), but also offers several degrees of freedom that have to be filled in by the school. For example, schools have to arrange placement situations where students can actually perform the Criticial Job Situation in its whole width, identity and train assessors, and develop transparent information systems for explaining this new way of assessing to participants and employers. The quality of the national assessment framework can only be derived from examining resulting *actual CBAs* in educational practice (Kane, 2008; Van der Vleuten & Schuwirth, 2005).

[1] VET trajectories in the Netherlands vary from in duration from one to four years. The number of summative CBAs depends on the length of the trajectory.

The two CBAs in this study covered two different competency profiles at two levels of VET education,[2] namely animal care specialist (*ACS*) at level 3 and assistant animal care specialist (*AACS*) at level 2. The levels mainly differed in that the level 3 incorporates more theoretical underpinning and thinking and a higher level of independence. The two actual CBAs were developed by two different teacher teams of one AVET school. By examining two CBAs within one school, the context variables disturbing the implementation of the national assessment into an actual assessment were held constant.

The CJS of the ACS assessment was titled 'take care dairy' which required students to independently work on a farm and take care of farm animals (core tasks: feeding, caring, milking, facilitating reproduction; competencies: e.g., collaborating with colleagues, applying professional knowledge). The CJS of the AACS was 'working with animals', which required students to take care of companion animals for example in an animal home or pet shop under supervision (core tasks are: feeding, handling, caring, playing; competencies: e.g. following instructions and procedures, using equipment and materials). During a period of 16 weeks students performed a number of activities in their own work placement context (e.g., a farm or a pet shop) to practice with performing the CJS. After these 16 weeks, the formative (i.e., learning) trajectory ended and students started performing comparable activities in the same work placement context for summative assessment purposes.

Participants

Four stakeholders groups were involved in this study. These were the developers of the national assessment framework ($n = 26$), representing teachers from all VET schools and five representatives of different fields of work; teachers in the roles of developer/assessor of an actual CBA (level 2: $n = 3$; level 3: $n = 3$); students ($n = 7$, women = 4, men = 3, mean age = 17; $n = 18$, women = 13, men = 5, mean age = 17.29), and employers of the students' work placement contexts in the role of assessor ($n = 7$; $n = 19$).

Instruments

Mixed-methods instruments, namely questionnaires and semi-structured group interviews, were used. Both instrument were grounded in new quality criteria for competence-based assessment, derived from Baartman et al. (2006). The twelve criteria were slightly adapted or split up a bit further to fit the summative assessment framework of this study, instead of Baartmans competence-based assessment *program* that consists of a combination of several formative and summative assessments (see Table 2).

The questionnaire contained 5-point Likert-scale items covering the twelve quality criteria in seventeen scales (3–5 items covering every criterion) and three open questions dealing with the positive

[2] VET in the Netherlands consists of four levels with level 1 being the lowest, most practical instead of theoretical oriented level of VET and level 4 being the highest, most elaborate and specialized VET level. In all levels, learning and working are intertwined on a regular basis.

Table 2
Description of the quality criteria as used in this study (based on Baartman et al., 2006).

Criterion	Short description
Acceptability	Degree to which all key stakeholders have confidence in the assessment's quality for assessing professional functioning
Authenticity	Degree of resemblance between the assessment (task, context, criteria) and professional practice
Cognitive complexity	Degree to which the assessment reflects the cognitive skills needed in professional practice and enables the judgment of these thinking processes
Efficiency	Degree to which the carrying out the assessment is feasible, compared to its benefits
Comparability	Degree to which assessment tasks, criteria an procedure are consistent for all students with respect to key features
Fairness	Degree to which the assessment allows the assessee to show all competencies and allows assessors to assess all the required competencies.
Fitness for competence-based purposes	Degree to which the assessment connects with the goals of CBE (a) focus on integration of knowledge, skills, and attitudes (b) focus on professional behavior (performances) (c) increasing the responsibility of the student in the assessment process
Meaningfulness	Degree to which the assessment is of significant value for all stakeholders with respect to future job and/or personal development
Reproducibility of results	Degree to which decisions made on the basis of the results of the assessment are independent of assessor or specific assessment situations. Therefore, multiple assessors, assessment tasks and situations should be combined
Transparency	Degree to which the assessment (goals, criteria, procedure, etc.) is clear and understood for all stakeholders
Alignment of instruction-learning-assessment	Degree to which the assessment (competencies, tasks, activities, criteria, etc.) are compatible with (a) instruction and learning in school (or at the institution) (b) learning and activities in work placements situations
Educational consequences	Degree to which the assessment a stimulates (a) reflection and personal development (b) generic competence development (c) motivation

and negative aspects of the CBA, and its fitness for assessing professional competence. The questionnaires were filled in by all developers, students, and employers. The teacher groups were too small for the quantitative data to have any value, therefore teachers did not fill in the questionnaires. The questionnaires were almost identical, except for a small number of questions that a certain stakeholder group had no information about (e.g., questions dealing with costs were left out of the student questionnaires). From the seventeen scales, all groups filled in sixteen. A crucial difference was that the developers' answers reflected the quality they *expected* of the actual CBAs to be developed based on the national framework, while the employers and students answers reflected their *experienced* quality of a specific actual CBA.

Semi-structured focus group interviews were conducted and audio-taped. The interview schedule was structured along the quality criteria. In the developers group, one interview was conducted with five teachers representatives of five agricultural fields. Per CBA, one interview was conducted with the teacher group, one with a random sample of students ($n = 3$ and 4), one with a random sample of employers ($n = 3$ and 4).

Analysis

One-sample *t*-tests for all quality criteria were calculated per group. In addition, one-way ANOVAs were computed comparing the group means per criterion. Games–Howell post hoc corrections were used to control for the variations in number of participants per group (Field, 2000). When the group mean scores for a criterion were significantly higher than the neutral score of 3 (p-value of .05) in the eyes of all stakeholders, the criterion was regarded as being of good quality. On the other hand, when a criterion was consistently scored as not significantly higher that 3, this was regarded as indicating a challenging criterion. Differences between mean scores of stakeholders might signal challenging quality aspects as well. In addition, comparing the developer group with the student and employer groups illuminated differences between expected quality and experienced quality.

Miles and Huberman (1994) method of cross-case comparison was used to analyze the qualitative data. Transcribed interview data and the qualitative questionnaire answers were meaningfully reduced to data about quality criteria or CBA characteristics (data reduction). Then, the data were organized into seven matrices (one per stakeholder group and per CBA) categorizing stakeholders' statements in top-down fashion into the twelve quality criteria (data display). Matrices displayed evaluative responses (positive or negative) with respect to the quality criteria as well as arguments supporting these responses. Comparing the matrices between stakeholder groups and both CBAs allowed for drawing conclusions about CBA quality and CBA characteristics that were argued to determine this experienced quality. Researcher interpretations were controlled for by using the member check procedure (Guba & Lincoln, 1989), asking all interviewed groups to check whether the reduced data accurately displayed the issues discussed in the interviews. A second researcher independently categorized the data along the quality criteria (inter-rater reliability of .77) and verified drawn conclusions made by the first researcher (Guba & Lincoln, 1989). Only in a small portion of the drawn conclusions, more elaborate discussion was needed to reach consensus.

Results

Research question 1 dealt with how various stakeholders valued the quality of the CBAs, in terms of the twelve quality criteria (Table 2). Table 3 shows that the stakeholder groups valued most quality criteria as significantly higher than the neutral value of 3.

All groups rated only 3 or less of the 16 scales as *not* significantly higher than 3, except for the level 2 AACS students who scored 6 out of the sixteen scales not significantly higher than 3. *Authenticity*, *fitness for assessing integration of knowledge, skills and attitudes*, and *alignment between work placement activities and the assessment* were even unanimously valued higher that 4. Challenging criteria turned out to be: *comparability*, *fitness for self-directiveness*, *alignment between school instruction and the assess-*

Table 3
Experienced quality on the twelve quality criteria of the two students groups, the two employer groups, and the developers.

		ACS students (n = 18) M (SD)	AACS students (n = 7) M (SD)	ACS employers (n = 19) M (SD)	AACS employers (n = 7) M (SD)	Developers (n = 26) M (SD)	ANOVAs F (p-Value)
1	Authenticity	4.30 (.43)**	4.25 (.62)*	4.58 (.44)**	4.17 (1.04)*	4.23** (.56)	
2	Cognitive complexity	4.10 (.54)**	4.28 (.65)**	4.32 (.45)**	3.86 (1.05)	4.06** (.75)	
3	Acceptance	4.06 (.93)**	4.86 (.38)**	4.21 (.73)**	4.07 (.98)*	3.73** (.69)	2.71 (.04) S2 > D
4	Efficiency	–	–	4.37 (.43)**	4.25 (.68)**	3.68** (.72)	5.50 (.008) E3 > D
5	Comparability	3.96 (1.01)**	4.00 (1.54)	–	–	3.38 (1.05)	
6	Fairness	3.71 (.63)**	4.33 (.76)**	4.46 (.62)**	3.95 (.87)*	3.72** (.74)	3.77 (.008) E3 > S3, D
7a	Fitness for assessing: integration of knowledge, skills and attitudes	4.00 (1.12)**	4.50 (.55)**	4.74 (.45)**	4.29 (1.11)*	4.36** (.76)	
7b	Professional behavior	3.92 (.58)**	3.93 (.70)*	4.62 (.43)**	4.13 (.82)*	4.00** (.67)	3.93 (.006) E3 > S3, D
7c	Self-directiveness	3.25 (.58)*	3.10 (.74)	2.61 (.96)	2.43 (1.27)	2.85 (.76)	
8	Meaningfulness	4.09 (.73)**	4.40 (.23)**	4.71 (.49)**	4.23 (1.16)*	3.88** (.58)	3.92 (.007) E3 > D
9	Reproducibility of results	3.97 (72)**	4.22 (.69)	4.32 (.60)**	4.76 (.37)**	3.64** (.84)	4.33 (.004) E2, E3 > D, E2 > S3
10	Transparency	3.64 (.81)**	4.50 (.50)**	4.50 (.44)**	4.79 (.39)**	4.02** (.81)	5.99 (.000) E2, E3 > S3, E2 > D
11a	Alignment between: school instruction and assessments	2.95 (.84)	3.38(.77)	–	–	4.03 (.91)**	6.76 (.003), D > S3
11b	Work placement activities and assessment	4.03 (.68)**	4.46 (.51)**	4.32 (.48)**	4.63 (.21)**	–	
12a	Stimulating: reflection and personal development	2.18 (.26)	2.48 (.54)	4.00 (1.16)**	4.86 (.38)**	3.93** (.75)	4.21 (.005) D, E2 > S3
12b	Development of generic competencies	4.21 (.70)**	4.33 (1.37)	4.27 (.52)**	4.10 (.60)**	3.99** (.79)	
12c	Motivation	3.51 (1.01)	4.50 (.90)**	3.92(1.05)**	4.21(1.15)*	4.13** (.49)	
	Not significantly > 3 (of 16)	3	6	1	2	2	

S3, student animal care specialist, VET level 3; S2, student assistant animal care specialist, VET level 2; E3, employers animal care specialist, VET level 3; E2, employer assistant animal care specialist, VET level 2; D, developers.
* $p > .05$.
** $p > .01$.

ment, and *stimulating reflection and personal development*. With respect to the first two criteria (comparability and fitness for self-directiveness), all groups were critical, while the latter two criteria (alignment and stimulation of reflection and personal development) were challenging because they were appreciated by developers and employers, but not by both student groups.

Differences between the stakeholders and between the two CBAs

In general, the employers were the most positive group, while the developers were the most negative (see right column Table 3). Student groups scored mostly in between. Significant differences showed that developers were more negative than one or both employer groups with respect to various quality criteria: *efficiency*, *fairness*, *fitness for assessing professional behavior*, *meaningfulness*, *reproducibility of results*, and *transparency*. On the other hand, developers were significantly more positive than level 3 students about *alignment between school instruction and assessment* and *stimulating reflection and development*.

Differences between both CBAs were negligible. No statistically significant differences were found between both employer groups or both student groups. In other words, the quality of the two actual CBAs developed based on the same national assessment framework were experienced to have comparable quality and quality problems.

Qualitative results: given arguments for experienced quality

Qualitative data gave insight into research questions 2 about what arguments stakeholders used to support their (quantitative) evaluative responses of the CBA. With respect to the highly valued quality aspects of the CBA, developers argued that because the national framework was developed in collaboration with and validated by the work field the assessment's *authenticity* and *alignment to work placement* were automatically warranted. Employers and students had more specific arguments for the assessment's *authenticity*, *integrative nature*, and its *alignment to work placement*: (a) directly observing student's performance of professional tasks in their work placement context; (b) involving the employer as co-assessor; (c) the holistic judgment focusing on ability to perform the job which is recognizable for employers, and (d) the use of multiple methods addressing professional competence in different ways. Employers stressed that not only the performance-on-the-job part made the CBA authentic, the CBI increased the assessments authenticity as well, as this addressed authentic professional thinking: "This CBI is asking all the questions that I (as a farmer) should actually be asking myself everyday" (ACS employer).

Arguments for the challenging quality criteria: (1) comparability

With respect to the challenging criteria, interview data showed an interesting pattern in the criterion *comparability*. Students and employers did not worry about incomparability as a result of the fact that all students performed their assessment at different farms or pet shops. They all agreed that the content of the assessment (i.e., the CJS, its core tasks and competencies) was comparable for all students, independent of placement context: "it does not matter if I have to milk the cows at this farm or at the next farm" (ACS student). However, all stakeholder groups doubted the comparability of the assessment procedure as used by different assessors. Developers and teachers doubted comparable use of assessment procedures, because of the newness of this way of assessing and lack of assessor training. Students and employers argued that the comparability was threatened for three reasons: (a) they expected some assessors to be stricter than others; (b) employers were unsure about their assessor role and doubted whether they would assess students in the same way as another employer, and (c) the relationship between student and employer, good or troubled, could blur the assessment procedure. On the other hand, employers stressed two characteristics of the national CBA framework that reduced the incomparability between assessors: first, combining an employer and a teacher assessor, and second, using a holistic overall judgment, focusing on the student's capability to perform in professional practice. This is a judgment that employers in the same field (e.g., different farmers) can

equally relate to and also stimulates them to be a critical assessor: "with this judgment I say that I would trust this student to take over my farm for a week, but also the farm of my neighbor. That is not something you just say. You have to be really sure" (ACS employer).

(2) Self-directiveness

With respect to fitness for *self-directiveness*, developers, teachers and employers agreed that the CBA was primarily teacher-guided. Developers and teachers argued that this had been a conscious choice, both in the national assessment framework as well as in the actual CBAs, at this time of experimenting with this new way of assessing students at the VET level. Student ratings were not very positive about this quality criterion, but the interviews showed that they did not experience this as a problem. They were not used to self-directiveness and were not searching for more responsibility: "Hmm. . . I did not think of that. I suppose I was given the opportunity to tell what I wanted to tell during the CBI" (ACS student). Thus, in this study, *increased student responsibility* is not seen as a crucial characteristic of CBA quality.

(3) Alignment and (4) educational consequences: stimulating reflection and development

Students did not appreciate the criteria alignment between school instruction and assessment and effectiveness for stimulating reflection and personal development. Developers expected that school curricula would properly prepare students for the CBA, while students complained about the lack of alignment between what they did in school and what they had to do in the CBA, supported by two arguments: (a) schoolwork consisted of discrete courses and a focus on theoretical knowledge, while the assessment required integration into performing a job task: "in school we do not work with the competencies that we are assessed on" (ACS student), and (b) students were not well-prepared for the kind of questions asked in the CBI. The CBI required them to deal with reflective knowledge or "why-questions" while students were used to learn and be asked for declarative knowledge or "what-questions": "I was surprised by the questions in the interview, I expected that the teacher would ask more knowledge questions", (AACS student). Employers also experienced that students were not well prepared for the CBI.

The quantitative finding that students did not experience the CBA to *stimulate reflection and personal development* was actually in line with the original intentions of the national assessment framework, in which the CBA was not to have this kind of formative purpose. In this respect, it was surprising that developers (who developed these guidelines themselves) expected the CBAs to stimulate reflection and development. Their arguments focused on the use of the CBI. They expected this interview to automatically stimulate reflection, which was corroborated by employer experiences "The CBI shakes-up the student. It forces him to be critical about his own actions and development" (ACS assessor). Students' responses suggested that they did not experience the CBI as such: merely having an interview for summative purposes does not automatically stimulate students to reflect and think about their development. These results suggest that the purposes of the CBA were not transparent, implicit, or at least open for multiple interpretations by various stakeholders.

Fairness and reproducibility of results: experienced preconditions

Quantitative data (Table 2) suggested that all stakeholders were positive about both the *fairness* and the *reproducibility of results* quality criteria, however qualitative data analysis necessitated a closer look at these two criteria. Fairness refers to whether the CBA allowed students to show all required competencies and allowed assessors to assess all these competencies. Reproducibility of results refers to whether the CBA allows for an accurate judgment of the student's competence, independent of assessor, assessment situations, or time. Employers, students, and teachers of the actual assessment argued that this CBA was *only* fair and reproducible when some preconditions were met: (a) the work placement context should allow the student to perform activities described in the CJS, which was not always the case; (b) the employer should be a co-assessor, a condition prescribed in the national framework, and (c) the summative CBA should not be treated as completely separate from students' activities during the preceding learning period. This is contrary to the guidelines of the national assessment framework that strictly separated learning and assessment activities. Instead, in the actual assessments, employers made use of: (1) activities performed during the placement period (i.e., the learning period of 16 weeks): "Only looking at the performances during the last week (i.e., the observed performance element of the CBA) is unfair, and unnatural. Something can go wrong, while he performed the task perfectly several times before. Everybody makes mistakes sometimes" (ACS employer). In addition, both employers and teachers also build their judgment on (2) a pre-conditional portfolio filled with assignments and tests that students had to satisfy *before* they were allowed to do the summative CBA. This pre-conditional portfolio was developed by the school. It was not an element in the national assessment framework. In other words, the pre-conditional portfolio was no official element of the summative CBA, but it was treated as such in practice: "I know that it is impossible to observe and discuss everything in the CBA, but that is not necessary, therefore I trust that pre-conditional portfolio already covered all separate competencies" (AACS teacher). Thus, contrary to the national assessment framework and the opinion of developers, stakeholders of the actual assessments agreed that for the CBA to be fair and reproducible it required taking into account additional information about students' performance over a longer period of time, next to the performance-on-the-job and the CBI.

Conclusion

Combining findings from two actual CBAs and triangulating data from various stakeholders allowed for identifying positive and challenging quality aspects of the national assessment framework for AVET that was build on theoretical notions of high quality CBAs. To explain their quality experiences, multiple stakeholders referred to many theoretical CBA characteristics (Table 1). Several theoretical characteristics were directly supported, while other where refined with specific characteristics necessary for assessment quality in actual educational practice. In the discussion section we will reflect on theoretical CBA characteristics in the light of the empirical findings. The findings also corroborated the quality of the national assessment framework in AVET education in the Netherlands, supporting national initiatives in changing towards competence-based assessment practices and setting guidelines for developing quality CBAs (Leigh et al., 2007). However, examining the quality of a CBA requires examination of the actual assessments as used in practice as shown by the differences between the expected quality of the developers and the experienced quality of the users. The arguments given by stakeholders of the actual assessment suggest that the national framework alone does not guarantee its quality. Various conditions have to be met in the actual use of the assessment in practice.

Discussion

Theoretical CBA characteristics in practice

This study provides *empirical* support from educational practice for several *theoretical* characteristics of CBAs (see Table 1). First, *integrating learning and assessment* activities and allowing the incorporation of a broad range of learning activities in summative assessments is previously emphasized to be positive for student learning (Birenbaum et al., 2006; Dochy, 2005; Harlen, 2005) and is in this study found to be important for fair and reproducible assessments. This study elaborates that in the case of workplace CBAs, as used in this study, integration between assessment and learning activities in school as well as between assessment and activities conducted in the work context is crucial. Second, stakeholders supported the characteristic of *combining multiple methods* and refined it by stressing that this mix of methods should: *incorporate a long-term measurement of student performance* (i.e., the placement period and/or the pre-conditional portfolio), an *actual observation of performance-on-the-job*, a *judgments from employers*, and *a method addressing authentic thinking processes* (i.e., the CBI). These characteristics were addressed in arguments for various quality criteria. Third, this study supports the importance of *collaborations between educational institutions and the work-field* in developing, conducting and evaluating quality CBAs of professional competence (Baker, 2007; Gulikers et al., 2007). However, where the developers seemed to feel that involving the employers in the validation of the assessments (i.e., only in the developing phase) guarantees authenticity of the assessment, the other stakeholders stress the necessity of involving them in the actual use of the assessments, for example as co-assessor. Fourth, an interesting refinement was made with respect to the *individualization* characteristic (Table 1), which determined the quality criterion of comparability. Individualization is favored with respect to the assessment context and specific content, but *standardization* in CBA should be guaranteed through assessment procedures that are equally used by all assessors. This also refers to fifth finding related to the quality criterion of *transparency*. The national assessment framework was supposed to lead to a transparent assessment system. However, several stakeholder arguments suggested that the transparency of the actual assessments needed improvement. Certainly when implementing a new assessment system, there should be more explicit communication about the roles and responsibilities of the (teacher and employer) assessor, about what is expected from students (e.g., ''why questions'' instead of ''what-questions''), and about the goal(s) of the assessment. Sixth, employers often referred to the characteristic of a *holistic overall judgment* on an assessment criterion directly related to job performance (e.g., would you trust this student to take over your farm for one week). This characteristic had a positive influence on many quality criteria in the eyes of the employers. This refines the theoretical characteristic of *criterion-reference scoring*: it argues against using a many criteria, but argues in favor of using criteria that directly address professional performance that employer assessors can directly relate to. This is an argument previously used as a crucial characteristic of authentic assessment (Gulikers et al., 2004). Several positive effects of judging holistically in summative assessments have been suggested before (e.g., Grainger et al., 2008). What this study adds in this respect is that a holistic judgment on crucial job-related criteria might also be an easy way to get the work-field more accepting of new CBAs and more involved for example as co-assessor.

One theoretical characteristic was not supported: *increased student responsibility and involvement*. In the national framework and in both actual CBAs evaluated in this study, this characteristic was not intended, not accomplished, but also not experienced as important for CBA quality. Previous studies argued that both teachers and students are not yet familiar with their changing roles in CBA in which more responsibility for the assessment should be transferred from the teacher to the students (Birenbaum et al., 2006; Biemans, Nieuwenhuis, Poell, Mulder, & Wesselink, 2004). Transferring responsibility and involvement to students cannot go without guidance or training. For example, for peer-assessment to work, students need to be trained in peer-assessment skills (Sluijsmans & Prins, 2006).

Separating or integrating formative and summative assessments?

In various arguments, this study showed the struggle between integrating or separating learning and assessment activities or formative and summative assessment purposes. This issue received a lot of recent attention in assessment research as well (e.g., Birenbaum et al., 2006; Harlen, 2005; Taras, 2005). The national assessment framework in this study prescribed a strict separation. This decision was guided by the requirements of the external quality assurance system for VET in the Netherlands. For a long time, the idea that formative and summative assessment should be strictly separated has been the dominant view in assessment research and practice (Black & Wiliam, 1998). This was expected to be required for guaranteeing assessment quality. In this study, however, stakeholders of the actual assessments experienced this strict separation to have a negative rather than a positive impact on CBA quality. Indeed, in conducting the actual assessments, stakeholders did not comply with the strict separation guideline of the national framework. Research is changing towards exploring ways in which formative and summative assessments can be clearly distinguished, but integrated in such a way that they support each other and lead to more effective and efficient assessment practices (Birenbaum et al., 2006; Harlen, 2005; Taras, 2005).

Differences between stakeholders: motivating and training teachers and developers

Contrary to other studies comparing (teacher) developer expectations with user experiences of assessment practices (e.g. Cummings & Maxwell, 1999; Gulikers, Kester, Kirschner, & Bastiaens, 2008; Maclellan, 2001), teacher developers in this study were more critical about many assessment aspects than the users, certainly the employers. A possible explanation is that the transition from traditional testing to competence-based practices requires a major shift for educational institutions and teachers (Biemans et al., 2004). This is fraught with uncertainties causing hesitation or skepticism about implementing CBAs on the part of the teachers. The results of this study suggest that several barriers *expected* by developers are not *experienced* as such by the users of the actual CBA. Of course, we should not loose sight of the fact that students and employers have a different perspective on and responsibilities in assessment practices and quality assurance than the educational institutions, however their experiences can play a vital role in motivating teachers in educational innovation processes (Gulikers et al., 2007).

Thoughts of caution

Interpreting and generalizing this study needs some thoughts of caution. First of all, this study deals with CBAs in the context of vocational education that prepares students for a concrete and clear future job. The CBA characteristics and operationalisations might look a bit different in other education levels like for examples secondary or university education. In addition, the

generalizability of these findings outside Dutch VET can be questioned. The Dutch government has a big say in what competence-based education in Dutch VET should look like and how its quality is to be determined. These governmental decisions are likely to create a reference frame for developing CBAs, evaluating their quality, and stakeholders' experiences (Johnston, 2004; Kane, 2008; Kaslow et al., 2007). Valid and meaningful examination of CBAs and their quality will always require taking the educational and political context into account (Kane, 2008; Slavin, 2008).

This study deals with stakeholders' *perceptions* of the quality of a CBA, being a subjective rating of its quality. It can be questioned to what extent these perceptions signal real, or objective, quality. However, perceptions *do* signal critical or strong characteristics of the CBA. The objective quality can be very high, if it is not perceived as such by the involved users, the CBA will never reach its intended results and quality (Gulikers et al., 2008; Van der Vleuten & Schuwirth, 2005). Also, the number of participants per group differed, some of which were relatively small. Even though corrections for group differences were dealt with in the analysis and where possible, this might have influenced the robustness of the findings.

Practical implications

Besides empirically supporting several theoretical notions of CBA quality, the findings can also be translated into practical guidelines for summative CBAs assessing professional competence in VET. These guidelines have to do with both the actual operationalisation of the assessment, but also with pre-conditional processes that should be taken into account.

1. Representatives of the work-field should be actively involved in the assessment process: as co-developer of the assessment to assure that the assessments validly reflect professional practice, but preferably also as co-assessor who has direct data about student's actual (and long-term) performance-on-the-job. However, the role and responsibilities of the employer in the RI should be clear, communicated and discussed, and understood by all.
2. A holistic overall judgment on criteria that directly relate to job performance can positively influence the involvement, acceptance and comparability of employers.
3. Individualization in assessment context and concrete content should be allowed, but standardization in assessment procedure and use thereof should be guaranteed.
4. The CBA should incorporate evidence of the student's actual performance (observed).
5. A summative CBA requires combining multiple methods that address the required competencies or job tasks from different angles. However, a fair and reproducible assessment: (a) requires the incorporation of long-term indications about the student's competence and performance (e.g., in a pre-conditional portfolio or long-term observation of practical performance); and (b) should allow involving a broad range of activities relevant for the competencies of the assessment, which implies no strict separation between activities conducted for learning and for assessment.
6. A summative CBA does not automatically have a formative effect on students. However, a summative CBA can have a formative function when the summative judgment is followed up (i.e., not tangled up) by good feedback and discussing this feedback in a dialogue with students

With respect to pre-conditions, this study suggests that for accepted summative CBAs:

7. A national assessment framework can set helpful guidelines, but still requires individual schools to contextualize and explicitly describe how the guidelines in the national framework are translated into an actual CBA in this specific educational context.
8. An explicit and elaborate description of the CBA goal(s), criteria, procedure and roles of all involved parties is required. This is needed for transparency and comparability between assessors. A shared understanding between stakeholders is also important for CBA quality in general.
9. A smooth alignment between (school/workplace) learning and assessment activities is pre-conditional. This also means assuring that students can perform and practice all required assessment activities in school and/or work placement context.

Overall, the change towards new competence-based assessment is a challenging one. A national acknowledged and collaborative approach, as was the case in this study, seemed to be a fruitful one (see also Kaslow et al., 2007; Leigh et al., 2007). However, evaluating actual assessment practices that schools implement based on this national intended assessment framework is needed to get more grip on what actually works and does not work in practice (Van der Vleuten & Schuwirth, 2005). By doing this, this study contributes to the knowledge-based about competence-based assessment and stimulates educational practice and future assessment research.

References

Baartman, L. K. J., Bastiaens, T. J., Kirschner, P. A., & van der Vleuten, C. P. M. (2006). The wheel of competency assessment: Presenting quality criteria for competency assessment programmes. *Studies in Educational Evaluation, 32*, 153–170.

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., Kirschner, P. A., & Vleuten, C. P. M. v. d. (2007b). Teachers' opinions on quality criteria for Competency Assessment Programs. *Teaching and Teacher Education, 23*(6), 857–867.

Baker, E. (2007). Presidential Address held at the annual Conference of the American Educational Research Association. Chicago, USA.

Benett, Y. (1993). The validity and reliability of assessments and self-assessments of work-based learning. *Assessment & Evaluation in Higher Education, 18*(2), 83–94.

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.

Birenbaum, M. (2007). Evaluating the assessment: Sources of evidence for quality assurance. *Studies in Educational Evaluation, 33*, 29–49.

Birenbaum, M., Breuer, K., Cascallar, E., Dochy, F., Dori, Y., Ridgeway, J., et al. (2006). A learning integrated assessment system. *Educational Research Review, 1*, 61–69.

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

Cummings, J. J., & Maxwell, G. S. (1999). Contextualising authentic assessment. *Assessment in Education: Principles, Policy & Practice, 6*, 177–194.

Dierick, S., & Dochy, F. (2001). New lines in edumetrics: New forms of assessment lead to new assessment criteria. *Studies in Educational Evaluation, 27*(4), 307–329.

Dochy, F. (2005, August). 'Learning lasting for life' and 'assessment': How far did we progress?. Presidential address EARLI 2005 at the 20th European Association for Research on Learning and Instruction, Nicosia, Cyprus.

Field, A. P. (2000). *Discovering statistics using SPSS for Windows: Advanced techniques for the beginner*. London: Sage.

Grainger, P., Purnell, K., & Zipf, R. (2008). Judging quality through substantive conversations between markers. *Assessment & Evaluation in Higher Education, 33*(2), 133–142.

Guba, E. G., & Lincoln, Y. S. (1989). *Fourth generation evaluation*. London: London Sage.

Gulikers, J., Bastiaens, T., & Kirschner, P. (2004). A five-dimensional framework for authentic assessment. *Educational Technology Research and Development, 52*(3), 67–85.

Gulikers, J. T. M., Bastiaens, T. J., Kirschner, P. A., & Kester, L. (2006). Relations between student perceptions of assessment authenticity, study approach and learning outcome. *Studies in Educational Evaluation, 32*, 381–400.

Gulikers, J., Biemans, H., & Mulder, M. (2007, September). *Evaluating the quality of competence-based assessment by involving multiple stakeholders*. Paper presented at the European Conference for Educational Research, Ghent, Belgium.

Gulikers, J. T. M., Kester, L., Kirschner, P. A., & Bastiaens, T. J. (2008). The effect of practical experience on perceptions of assessment authenticity, study approach, and learning outcomes. *Learning and Instruction, 18*, 172–186.

Harlen, W. (2005). Teachers' summative practices and assessment for learning—Tensions and synergies. *The Curriculum Journal, 16*(2), 207–223.

Johnston, B. (2004). Summative assessment of portfolios: an examination of different approaches to agreement over outcomes. *Studies in Higher Education, 29*(3), 395–412.

Kane, M. T. (2008). Terminology, emphasis, and utility in validity. *Educational Researcher, 37*(2), 76–82.

Kaslow, N. J., Rubin, N. J., Bebau, M. J., Leigh, I. W., Lichtenberg, J. W., Nelson, P. D., et al. (2007). Guiding principles and recommendations for assessment of competence. *Professional Psychology: Research and Practice, 38*, 441–451.

Leigh, I. W., Smith, I. L., Bebeau, M. J., Lichtenberg, J. W., Nelson, P. D., Portnoy, S., et al. (2007). Competency assessment models. *Professional Psychology: Research and Practice, 38*(5), 463–473.

Linn, R. L., Baker, E. L., & Dunbar, S. B. (1991). Complex, performance-based assessment: Expectations and validation criteria. *Educational Researcher, 20*(8), 15–21.

Maclellan, E. (2001). Assessment for learning: The differing perceptions of tutors and students. *Assessment and Evaluation in Higher Education, 26*(4), 307–318.

Messick, S. (1994). The interplay of evidence and consequences in the validation of performance assessments. *Educational Researcher, 23*(2), 13–23.

Miles, M. B., & Huberman, A. M. (1994). *Qualitative data analysis. An expanded sourcebook*. Thousand Oaks: Sage Publications.

Schön, D. (1987). *Educating the reflective practitioner*. San Francisco: Jossey-Bass.

Schuwirth, L. W. T., & van der Vleuten, C. P. M. (2006). A plea for new psychometric models in educational assessment. *Medical Education, 40*, 296–300.

Segers, M., & Dochy, F. (2006). Enhancing student learning through assessment: Alignment between levels of assessment and different effects on learning. *Studies In Educational Evaluation, 32*(3), 171–179.

Segers, M., Dochy, F., & Cascallar, E. (2003). *Optimising new modes of assessment: In search of qualities and standards*. Dordrecht: Kluwer Academic Press.

Slavin, R. E. (2008). Perspectives on evidence-based research in education: What works? Issues in synthesizing educational program evaluations. *Educational Researcher, 37*, 5–14.

Sluijsmans, D., & Prins, F. (2006). A conceptual framework for integrating peer assessment in teacher education. *Studies in Educational Evaluation, 32*, 6–22.

Smith, K. (2007). Empowering school- and university-based teacher educators as assessors: A school–university cooperation. *Educational Research and Evaluation, 13*(3), 279–293.

Strickland, A., Simons, M., Harris, R., Robertson, I., & Harford, M. (2001). On- and off-job approaches to learning and assessment in apprenticeships and traineeships. In N. Smart (Ed.), *Australian Apprenticeships: research findings* (pp. 199–220). Leabrook: National Centre for Vocational Education Research Ltd.

Struyven, K., Dochy, F., & Janssens, S. (2003). Students' perceptions about new modes of assessment in higher education: A review. *Assessment & Evaluation in Higher Education, 30*(4), 331–347.

Taras, M. (2005). Assessment – summative and formative – some theoretical reflections. *British Journal of Educational Studies, 53*(4), 466–478.

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