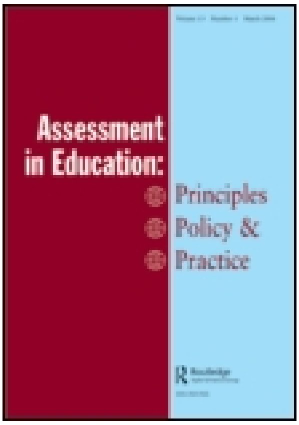
The explication of quality standards in self-evaluation

DOI:
10.1080/0969594X.2011.570731

Document status and date:
Published: 01/01/2012

Document Version:
Publisher's PDF, also known as Version of Record (includes final page, issue and volume numbers)



Assessment in Education: Principles, Policy & Practice

Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/caie20

The explication of quality standards in self-evaluation

Larike H. Bronkhorst [a], Liesbeth K.J. Baartman [b] & Karel M. Stokking [a]

[a] Department of Education, Faculty of Social and Behavioural Sciences, Utrecht University, Utrecht, The Netherlands

[b] Eindhoven School of Education, Eindhoven University of Technology, Eindhoven, The Netherlands

Version of record first published: 19 Dec 2011

To cite this article: Larike H. Bronkhorst, Liesbeth K.J. Baartman & Karel M. Stokking (2012): The explication of quality standards in self-evaluation, Assessment in Education: Principles, Policy & Practice, 19:3, 357-378

To link to this article: http://dx.doi.org/10.1080/0969594X.2011.570731

Assessment in Education: Principles, Policy & Practice
Vol. 19. No. 3, August 2012, 357–378

The explication of quality standards in self-evaluation

Larike H. Bronkhorst[a]*, Liesbeth K.J. Baartman[b] and Karel M. Stokking[a]

[a]Department of Education, Faculty of Social and Behavioural Sciences, Utrecht University, Utrecht, The Netherlands; [b]Eindhoven School of Education, Eindhoven University of Technology, Eindhoven, The Netherlands

Education aiming at students' competence development asks for new assessment methods. The quality of these methods needs to be assured using adapted quality criteria and accompanying standards. As such standards are not widely available, this study sets out to examine what level of compliance with quality criteria stakeholders consider satisfactory. Two professional education programmes specified the implicit standards they applied in a self-evaluation procedure designed to evaluate the quality of their Competence Assessment Programs (CAPs). They specified similar cut-off scores, but different descriptive standards. Analysis revealed that this was due to their experience with competence-based education and the quality of their own CAP, but influences of the selected method and the understanding of the quality criteria were also found. As such, the specified standards are local, but meaningful for the programmes' quality assurance. Implications for self-evaluation and standard-setting procedures are discussed.

Keywords: assessment; quality assurance; standards; self-evaluation

With the introduction of competence-based education a wide variety of new assessment methods have been developed, all directed at adequately assessing competence development. Many educational institutions developed and implemented (entirely) new assessment methods, such as portfolios or performance assessments, for which neither quality criteria nor quality standards were at hand. As assessment is a powerful determinant of learning, its quality should be guaranteed. Yet concerns have been expressed about the quality of these new assessment methods and the ways of establishing their quality (Birenbaum 2007). Increasingly, self-evaluations are used to evaluate educational quality (McNamara and O'Hara 2008), from which a number of challenges have emerged. One of these is the fact that the frame of reference used appears to have a considerable effect on the outcome of the evaluation (Baartman et al. 2011). That is, the quality evaluation depends on the implicit standards to which the current situation is compared. This may be due to the fact that no external standards exist, but could also be the result of other factors (Berk 1996; Delandshere 2001). This raises the issue of what standards educational institutions use in quality assurance when no standards are available, and which factors are of influence in the process of determining what level of compliance is considered

*Corresponding author. Email: l.h.bronkhorst@uu.nl

ISSN 0969-594X print/ISSN 1465-329X online

http://dx.doi.org/10.1080/0969594X.2011.570731
http://www.tandfonline.com

satisfactory. This article aims to shed light on those questions using a cross-case analysis of the quality self-evaluations of two higher education programmes struggling to develop high quality assessments in an initial stage of developing competence-based assessments.

New notions about assessment quality

In response to societal changes, education is being increasingly oriented towards students' competence development. Competence can be defined as an integration of the knowledge, skills, and attitudes required to adequately function in a professional environment (Lizzio and Wilson 2004), in which the competence to acquire knowledge is more important than the possession of knowledge itself (Dochy and McDowell 1997). By introducing competences as the ultimate goal of education, the gap between education and profession is assumed to be reduced. This perspective on education parallels socio-constructivist theories of learning (Tynjälä 1999; Birenbaum 2003), which emphasise the idea of knowledge as context-dependent, requiring meaningful learning activities and application in a realistic context. Since assessment profoundly influences learning (Biggs 1996), assessment has also undergone changes. Assessment has shifted from a culture of testing towards a culture of assessment (Dochy and McDowell 1997). This change has several implications. First of all, the content or objective of the assessment has shifted from separate knowledge and skills to integrated competences. Secondly, in an assessment culture, assessment is interconnected with the learning process (Wolf et al. 1991) as it aims both to measure the preceding as well as stimulate the subsequent learning. Thirdly, the assessment culture values an authentic context over a decontextualised event. Lastly, in an assessment culture the learner becomes more than the object of assessment. As assessing is believed to be beneficial for learning, learners are increasingly involved in the assessment process as self- or peer-assessors (Dochy 2001). All in all, assessment from this perspective can be seen as a process, instrument or method to guide and evaluate learner development.

These changes gave rise to diversity in assessment practices aimed at capturing not only knowledge, but also skills and attitudes, as well as their integration. Consequently, single assessments are no longer suitable for determining competence development and/or proficiency (Van der Vleuten and Schuwirth 2005; Knight 2000). Instead, several competence indicators (i.e. assessment methods) need to be used in a coherent whole (Birenbaum 1996), which Baartman and colleagues (2006) have termed a Competence Assessment Programme (CAP). In such a CAP different assessment methods are combined, including both knowledge tests such as multiple choice or open questioning, and newly developed methods such as portfolios, criterion-based interviews and performance assessments. Assessments in a CAP can have both formative and summative functions depending on the context and the goals of the educational programme.

Along with the development of new assessment methods, discussions have arisen about appropriate criteria and procedures to judge their quality. Concerns have been expressed about the quality of these new assessment methods and the ways of establishing this quality (Birenbaum 2007). Some authors have proposed new quality criteria (e.g. Driessen et al. 2005) or a widened set of quality criteria to do justice to the fact that these assessments have additional purposes besides assessing existing knowledge and skills (Linn, Baker, and Dunbar 1991; Gielen, Dochy, and Dierick 2003), such as stimulating subsequent learning. Others argue that the criteria of validity and reliability should be adapted to these often more qualitative and open-ended assessments (Bennett 1993). Next to that, the procedures of assessment quality assurance have also been subjected to change. Analogous to the new roles students play in their assessment, there has been a trend towards self-evaluation of education and assessment practices (McNamara and O'Hara 2008; Nevo 2001) as a way to increase educational institutes' empowerment, ownership and professionalisation. Self- or internal evaluation is carried out by an educational institution itself, for example by a group of teachers, the department or school manager, a specific staff member, or a combination thereof (Baartman et al. 2007). In contrast, external evaluation is carried out by someone outside the school, usually inspectors or governmental organisations, and mainly serves accountability purposes (Nevo 1994).

As argued before, widely accepted criteria for evaluating CAP quality do not exist, let alone standards. As the goal of this article is to study the standards used by educational institutions, in the absence of national or external standards, we do not further discuss the appropriateness of the different sets of quality criteria here. For this study, the 12 quality criteria developed by Baartman and colleagues (2006) were chosen, as these provide a combination of both the well-established criteria of validity and reliability, adapted for new assessments, complemented by new quality criteria to do justice to the often formative nature of new assessments. As similar combinations of quality criteria have been proposed by other authors, the results of this study could be generalised to these sets of criteria. The quality criteria used are depicted in Table 1.

An evaluation is always a comparative process which requires a frame of reference (Sadler 1989, 1998). Accordingly, any appraisal, be it the assessment of student competence or the evaluation of a CAP, must be characterised by a clear

Table 1. Quality criteria.

Quality criterion	Short description
Fitness for purpose	Alignment among curriculum and CAP.
Reproducibility of decisions	The combination of multiple assessors, assessment tasks and assessment situations.
Transparency	CAPs should be clear and understandable to all stakeholders (e.g. students, teachers, employers).
Acceptability	All stakeholders should approve of the assessment criteria and the way the CAP is carried out.
Comparability	The tasks, criteria, working conditions and assessment procedures should be consistent with respect to key features of interest.
Fairness	Students should get a fair chance to demonstrate their competences.
Fitness for self-assessment	CAPs should stimulate self-regulated learning.
Meaningfulness	CAPs should have a significant value for all stakeholders involved.
Authenticity	The degree of resemblance of a CAP to the future workplace.
Cognitive complexity	A CAP should enable the judgement of thinking processes (and application of knowledge).
Educational consequences	The degree to which the CAP and its results yield positive effects on learning and instruction.
Costs and efficiency	The feasibility of carrying out the CAP for assessors and students.

notion of the relevant criteria and the accompanying standards. For the old quality criteria (i.e., reliability and validity) such standards were often specified in terms of a cut-off score. For instance, the test-retest reliability should be at least .70 (e.g. Downing 2003). For the new quality criteria, no such scores or other standards have been developed, while the desired level of attainment on each of the quality criteria (i.e. the standards) is indispensable as a frame of reference in quality assurance.

Specifying quality standards

In general there are various ways of *specifying* a standard. Sadler (1987), discussing student assessment, distinguished four ways to specify educational standards. Typically, standards were made explicit by establishing *numerical cut-off* scores specifying the minimum qualifying level, and possibly also other levels of attainment. Such scores are called 'sharp standards' and their utilisation is rather straightforward. However, prolonged agreement on such scores is difficult to accomplish as is an unequivocal interpretation of what the scores entail. A second way to specify a standard is to rely on *tacit knowledge* or connoisseurship of the assessors. Here, the standards reside solely in assessors' heads and a derivative of the standard becomes apparent only after an evaluation. Although many assessment practices rely on such policies, such tacit standards are criticised for being non-transparent, inconsistent, labour-intensive and based on fragile argumentation for consensus (cf. Purves 1993). Two more powerful ways (Sadler 1987) of specifying standards for formative purposes are using exemplars and verbal descriptions. *Exemplars* are key examples chosen to designate desired levels of proficiency. As this yields a highly concrete standard, it is most appropriate for product assessment, although it can also be used for other purposes. However, collecting exemplars for a relatively large number of criteria creates logistical problems. Specifying standards in terms of *verbal descriptions* or 'qualitative rubrics' (Scriven 1980) entails denoting the properties that characterise the desired level of quality. Such standards have been developed in response to tacit assessment and are a way to objectify standards and make them publicly accessible. Yet such standards tend to be somewhat fuzzy as they are given in linguistic terms. Due to the fact that both the criterion for which a standard should be specified and the specified level (such as 'mediocre') of that specific standard require interpretation (Koretz and Deibert 1995), these standards depend on the context and cannot be understood, or applied, without this context. Nevertheless, Sadler argues that verbal descriptions are often the most feasible to use in assessments, especially those evaluations which require multiple criteria. As these criteria are often inter-related, they can be complementary, but also conflicting. Therefore, potential trade-offs between the criteria should be considered (Haladyna and Hess 1999), which verbal descriptions may afford.

Standard setting

Next to the diversity in the ways in which standards can be specified, there are many procedures designed to settle on the desired level of attainment (i.e. ways to set the standard). Berk (1996) presented nearly 50 standard-setting procedures designed for different standard-setting applications. The choice for a specific standard-setting procedure should be guided by the type of product being evaluated and the number of criteria invoked in the process (Hambleton et al. 2000). In the case of a CAP quality evaluation, as in the current study, the product to be evaluated, a CAP, contains various assessment methods, and a widened set of quality criteria (Gielen, Dochy, and Dierick 2003) should be taken into account in its evaluation. This creates complex conditions for which existing standard-setting procedures are not adequately equipped.

Equivocal understanding of criteria has been found to influence the outcomes of more traditional standard-setting procedures (Skorupski and Hambleton 2005). Price (2005) and Sadler (1987) argue that standards resulting from standard-setting procedures are contingent on the local situation and the standard-setting procedure selected, as all procedures rely on human judgement (Norcini and Shea 1997; Berk 1996). Accordingly, in the literature many examples can be found of difficulties with replicating standard-setting procedures, casting doubt on their objectivity (Hambleton et al. 2000).

On the other hand, specifying standards in self-evaluations may come down to making the expectations of the stakeholders explicit. Vanhoof and Van Petegem argue that 'quality assurance implies having an idea of what quality involves' (2007, 102). They claim that such ideas are often expressed in expectations. These expectations designate the desired level of quality and can be *school-external* and *school-internal*. External expectations comprise the external requirements for accountability and accreditation (e.g. national standards), whereas internal expectations consist of the aspects that the educational institution itself considers crucial for educational quality. Both types of expectations can be considered quality standards and can be used in quality evaluation, but they serve a different purpose (i.e. accountability vs. improvement). Internal evaluations are aimed at the improvement of educational quality (Nevo 2001), typically using self-evaluation, an increasingly common practice of school evaluation (McNamara and O'Hara 2008). CAP quality self-evaluation could make use of the internal expectations of the educational institution's stakeholders. Yet the stakeholders expectations are often implicit (Price 2005), whereas to be of use in self-evaluation, they should be made explicit.

Resulting from these considerations, the CAP quality self-evaluation method developed by Baartman and colleagues (2007) was adapted in the current study to incorporate the explication of the standards. To that end, we incorporated a group discussion during the procedure. This is advanced as an adequate way to explicate implicit standards residing in the heads of assessors (O'Donovan, Price, and Rust 2004). A discussion is also an appropriate way to establish the necessary intersubjectivity among the evaluators and to reach consensus on the desired level of attainment, which makes future formative evaluation more effective and powerful (Vanhoof and Van Petegem 2007). Results of research using this self-evaluation procedure (Baartman et al. 2011) indicated that during a CAP quality self-evaluation, the participants compared their actual CAP to some standard, without specifying it. The authors argued that schools operate from different frames of reference, use the quality criteria in different ways and give different examples to account for (the same) perceived quality of their CAPs. These results are already an indication that CAP quality criteria and accompanying (implicit) standards are not unequivocal, but open to multiple interpretations.

In recognising that standard-setting procedures may not lead to objective, general standards, an essential question becomes what influences the results of these procedures (McGinty 2005), and to explore how this might affect quality assurance. To shed light on this question, we choose to compare the self-evaluation and specified standards by stakeholders involved in similar CAPs (Norcini and Shea 1997) using the same procedure and the 12 quality criteria as the conceptual framework for comparison (Greene and David 1984). The specific research questions that guide the current study are: (1) What implicit standards do educational programmes apply in a quality self-evaluation procedure? (2) What are the differences in how the standards are applied? (3) How can these differences be explained? A cross-case comparison was carried out to explore these questions.

Method

Participating institutions

The current study was carried out at a large university of professional education in the Netherlands, a type of education comparable to polytechnics. The institute offers four-year bachelor programmes with a strong vocational emphasis (Van Berkel and Wolfhagen 2002). In response to the societal changes described earlier, most education institutes of this type have introduced competence-based education.

We selected two programmes to participate in this study, which should be seen as contrasting cases. One was an Applied Natural Sciences (ANS) training programme and the other a teacher training (TT) programme. These two programmes were selected so as to differ only marginally in their CAP characteristics (following Norcini and Shea 1997), as they were part of the same educational institute and evaluated the CAP of the same year. At the same time the two programmes contain variation on potential key explanatory variables (Greene and David 1984) such as professional orientation and experience with competence-based education. As a strong link between education and the profession is believed to be essential in competence-based education (Tillema, Kessels, and Meijers 2000), dissimilar professional orientations might induce differences in the understanding of the quality criteria and/or importance assigned to the quality criteria. For instance, fitness for purpose might be conceived differently when perceived from a programme with a clear professional orientation – like the TT programme – than in a broad ANS programme based on both chemical and biological aspects. In terms of experiences with competence-based education and assessment, the TT programme had less experience than the ANS programme, which was considered to be a forerunner within the university in terms of its assessment practices. Therefore the ANS standards were expected to be at least different, but perhaps also more in line with new notions about competence assessment than the TT standards.

Participants

As participant selection is considered crucial in standard-setting as well as self-evaluation procedures, an informant from the education institute was asked to select seven different stakeholders from each of the two programmes based on the following criteria: (1) participants needed to have extensive knowledge of the CAP employed (Berk 1996; Norcini and Shea 1997) as well as personal involvement (Vanhoof and Van Petegem 2007); (2) there needed to be a clear difference in stakeholders' practical experience with the CAP and their influence on the policies and regulations of the CAP (following Berk). In practice this meant that, for instance, a principal concerned mostly with policy and a teacher with practical experience both contributed to the group discussion. These diverging backgrounds not only enable different perspectives on the actual CAP and thereby a more profound discussion on its quality, but also present different perspectives on what is important in assessment (McGinty 2005), possibly yielding different standards.

Procedure

Training session

All participants attended an explanatory training session. The purpose of this session was threefold. Firstly, past research showed that participants from the same educational programme differed in their knowledge and therefore their definition of their CAP (Baartman et al. 2007). For that reason a part of the training session was dedicated to the joint establishment of what the programme's CAP exactly entailed, enabling an analysis of the influence of the local situation in terms of the implemented CAP. Secondly, the participants needed a clear and correct understanding of the quality criteria (following Sadler 1998), as criteria used in many standard-setting studies can be open to multiple interpretations (Skorupski and Hambleton 2005). As we considered an adequate understanding of the quality criteria to be a prerequisite for the evaluation, the remainder of the session was devoted to the explanation of the CAP framework and its 12 quality criteria. The participants' understanding of the quality criteria was assessed at the end of the session with a multiple choice test developed by the authors. The test was graded by the first two authors on a scale from 0 to 10. The ANS programme scored an average of 7.56 (SD = 1.16) and the TT programme scored an average 7.36 (SD = 1.46), which indicated that the participants had a sufficient comprehension of the quality criteria.

Thirdly, the participants were walked through the subsequent procedure. The participants were asked to participate in the subsequent steps of the procedure according to their stake in the assessment for which they had been selected (i.e. the second participant selection criterion). Next to that, the operationalisation of the concept standard for current purposes was discussed. In other studies on specifying standards, participants were required to specify different levels of proficiency, such as 'insufficient', 'poor', 'sufficient', or 'good', creating a rubric with scaled levels of achievement (Allen and Tanner 2006). Although such a complete rubric has benefits in terms of transparency, the current study already contains a high degree of difficulty due to the complexity of a CAP and the large number of criteria for which standards should be specified. To avoid further cognitive demands, we chose to focus on one level of proficiency, namely 'satisfactory'. This was defined as the level of compliance the participants would be satisfied with, according to their stake in the assessment. Hence, this level is higher than a conventional cut-off score, denoting minimally sufficient quality, but lower than the participants' ideal. This level was chosen, as a description of satisfactory quality can function as an exemplar to guide subsequent improvements, while still being realistic.

Individual evaluation questionnaire

After this training session, each participant had a week to individually fill out a CAP quality evaluation questionnaire available on the internet. The 12 quality criteria were operationalised in the form of indicators; concrete aspects of a quality criterion in practice (for a more elaborate description, see Baartman et al. 2007). Participants were asked to score their actual CAP on each indicator and indicate whether they considered this score to be satisfactory, as an indication for the (implicit) standards they use.

These scores could range from 0 to 100, but this was invisible to the participants as they positioned an analogue slide bar between 'not at all' and 'completely' to avoid the impression of grading. Participants were invited to present argumentation, examples or a rationale for their rating of the actual CAP and for its sufficiency. After this, the participants ranked the quality criteria on the basis of their importance by first dividing them equally into two categories (most and least important) and then arranging the criteria in each category in descending order of importance. This ranking was required, because when multiple criteria are invoked in evaluation, participants need to make mental comparisons and consider trade-offs between the criteria (Sadler 1987). We expected the criteria regarded as most important to receive priority in such a comparison, which might influence the standards.

Semi-structured group interview

Based on information from both the training and the web-based questionnaire, a semi-structured group interview was held to have the participants discuss personal and group findings, using both open and probing questions. Probing questions concerned individual differences in the scores and in the sufficiency ratings that the participants had given in the web-based questionnaire. The two open questions that were asked for every quality criterion focused on the extent to which the actual CAP complied with that particular criterion (i.e. the actual situation) and which level of attainment the participants considered to be satisfactory (i.e. the standard). The latter question had to be answered by the participants by giving exemplars or verbal descriptions of a CAP with satisfactory quality on that particular criterion. As stated previously, participants were asked to specify a single level of compliance or standard per criterion.

During the group interview, the participants discussed the 12 quality criteria one by one. Before moving to the next quality criterion, participants were asked to indicate their evaluation of the actual situation and their standard on the criterion just discussed. To avoid numerical associations and the impression of grading, this was done by having them mark their responses to the questions 'To what extent does your CAP comply with [quality criterion]' and 'What extent would be satisfactory?', each on a line between 'not at all' and 'completely'. Following the methodology described by O'Donovan and colleagues (2004), after the group discussion a well-reasoned judgement on a criterion level was expected.

After the interview the participants were asked to individually determine the level of difficulty of the training, the CAP quality evaluation questionnaire, the group interview and the procedure as a whole to obtain an indication of their understanding (Van Der Schaaf, Stokking, and Verloop 2003). As limited understanding might have an erroneous effect on the results, this possibility should be ruled out. The results indicated that the procedure had not been too difficult. The group interview was considered the most difficult part (mean ANS = 6.1, SD = 2.4; mean TT = 4.6, SD = 1.6 on a scale from 0 to 10 (0 = not difficult, 10 = very difficult)). The interviews lasted two and a half hours. All interviews were video-taped with the permission of the participants.

Data analysis

To answer the first research question, the group interviews were transcribed verbatim followed by qualitative content analysis. First of all, although the quality criteria had been discussed sequentially, some topics overlapped and protocols needed to be arranged thematically. The segmentation according to the 12 quality criteria was done using MEPA (Erkens 2005). Two researchers coded independently and the inter-rater reliability was high (Cohen's κ = .87). Secondly, as participants discussed both the actual situation and their standards, the information per criterion was analysed further, separating the evaluation of the actual CAP from the specification of the standard. This was done by distinguishing participant contributions in terms of: (1) an evaluation of the actual CAP judged as containing satisfactory quality; (2) an evaluation of the actual CAP judged as containing unsatisfactory quality (without specifying an alternative); and (3) the standard descriptions. After a first round of independent coding, inter-rater reliability was mediocre (Cohen's κ = .47). The main differences pertained to the coding of comments about unsatisfactory actual quality when accompanied by an alternative that specified the standard. Once the researchers had discussed their differences, the inter-rater reliability became acceptable (Cohen's κ = .74). Based on this analysis, a coherent description of the actual CAP and its quality as seen by the participants as well as descriptive quality standards could be distilled. These standards were organised as a qualitative rubric with only one level (i.e. satisfactory), identifying the desired characteristics of the CAP in comprehensive, descriptive terms, making the standards most useful for subsequent evaluation (Allen and Tanner 2006).

To answer the second research question, these qualitative standards were complemented with the averaged cut-off scores the participants gave during the group interview. Subsequently, the two programmes were systematically compared using the quality criteria as a conceptual framework (Greene and David 1984). Differences between the cases in terms of their descriptive and numerical standards were documented. The first author did this independently, but her findings were checked by means of a condensed audit trail (Akkerman et al. 2008), in which the second author checked and verified the findings of the analyses. Small differences were found, resulting in minor alterations in accordance with both authors' opinions.

To answer the third research question, possible explanations of the differences between the standards were identified by examining all available data. Participants' ranking of the quality criteria in order of importance was given a score from 12 (most important) to one (least important). The scores were averaged across programmes, resulting in an average importance score for all criteria, which were combined with the CAP descriptions resulting from earlier analyses. Although this arrangement of information is not equal to that of a comprehensive case study, it did enable us to provide a meaningful rationale for the differences in the applied standards (Lichtman 2006). This rationale was put to the test by converting it into hypotheses, which were tested by re-examining the available data, looking for confirming and disconfirming evidence. Including the two key explanatory variables that had been introduced in the design, namely experience with competence-based education and professional orientation, and the hypothesis following from the ranking of importance, this resulted in seven hypotheses explaining the standard differences across programmes. The findings of the first author using this procedure were again verified by means of a condensed audit trail by the second author. Confirmed hypotheses are mentioned in the results section.

Results

CAP characteristics

To compare standards, the product for which they are specified (i.e. the CAP) should be similar (Norcini and Shea 1997). Even though the educational programmes were selected on the similarity of their CAPs, as they had the same CAP on paper, the CAP characteristics of the programmes differed in practice. As this had an impact on the resulting standards, the similarities and differences of the implemented CAPs of the two programmes are shortly described to be able to account for the CAP differences in answering the research questions.

Both CAPs were arranged around a student portfolio, in which students collected and presented various proofs of competence they had gathered during the year. These proofs of competence included, among other things, knowledge and skills tests, performance assessments, and self- and peer assessments. The portfolios were graded at the end of the academic year by two summative assessors. Several differences between the programmes in the implementation of these procedures are worth noting. Firstly, the ANS portfolio was graded solely in its written form. In contrast, a criterion-based interview was included in the overall policy of the university and the TT programme had incorporated an interview in which the students were questioned about their portfolio. Secondly, the ANS organised practice trials of all main assessments of its CAP to familiarise students with assessment procedures, which the TT programme had not, or had done only occasionally. Thirdly, the TT programme explicitly stated what should be in the portfolio. Among other competence proofs, students were obliged to include proofs they had selected and generated themselves. In contrast, the ANS programme had not specified the compulsory contents of the portfolio, and students were free to provide any proof they considered appropriate. In practice, ANS students could gather all the necessary proofs to show their competence by participating in the regular assessments. Finally, the TT programme made use of both internal assessors and external assessors of the professional field to reach a summative decision, whereas ANS only employed assessors from its own staff.

Standards descriptions

To answer the first research question the numerical cut-offs are listed in Table 2. The descriptive standards – in terms of exemplars and verbal descriptions – are specified in Appendix 1. Both programmes set the cut-off scores at about 80 for all the criteria, indicating that they required their CAP to comply to a large extent with all 12 quality criteria.

The stakeholder agreement for the cut-off scores was .01 (Cohen's κ) for ANS and .78 for TT. This indicated that the ANS stakeholders had very diverging perspectives on the level of compliance with the criteria. As the participants were selected on diversity, these results should not come as a surprise (Kane 1994) and they should not impede further interpretation of the results (McGinty 2005).

Table 2. Importance rating of, and actual and desired compliance with, the quality criteria.

	Applied Natural Sciences			Teacher Training		
Quality criterion	I[a]	A[b]	S[c]	I	A	S
Fitness for purpose	7.71	79	83	11.50	50	86
Reproducibility of decisions	6.71	89	86	5.50	64	82
Transparency	8.43	77	82	7.33	67	83
Acceptability	6.0	86	82	6.00	77	82
Comparability	8.71	82	82	6.50	56	76
Fairness	7.29	84	86	6.17	73	85
Fitness for self-assessment	5.14	89	83	7.00	60	82
Meaningfulness	5.29	86	83	7.33	74	78
Cognitive complexity	5.57	84	72	7.83	67	59
Authenticity	7.71	93	87	5.83	82	86
Educational consequences	3.00	82	76	3.83	73	69
Costs and efficiency	6.43	65	76	3.17	37	65

Notes: [a]Importance assigned to quality criterion (1–12); [b]Actual CAP rating (0–100); [c]Standard specified in terms of a cut-off score (0–100).

Standards differences

Although the cut-off scores are very similar, the descriptive standards specified by the two programmes contain various differences. The main differences lie in the way in which the programmes operationalise a satisfactory level of attainment. Firstly, the standards specified by ANS are more detailed than those specified by TT. Secondly, the ANS standards specify unconditional levels of compliance with the quality criteria, whereas many TT standards include conditions that have to be met by the educational programme *before* the actual standard can be met. Lastly, next to these general differences, there are also some instances in which the standards contained more fundamental differences. These differences are visible in Table 3. The accompanying quotes from the group interview illustrate some of the rationales that underlie these differences.

Quality standard rationales

To account for potential standard differences, the experience with competence-based education and the professional orientation guided the selection of the cases as possible explanatory factors. A measurement of importance assigned to the quality criteria was included for the same purpose. First of all, ANS' greater experience with competence-based education seemed to result in a more comprehensible vision of CAP quality, which enabled them to specify detailed standards supported by a rationale. As the TT stakeholders lacked such experience, their standards did not show the same amount of detail or contemplation. This difference in experience also caused a different attitude towards the various stakeholders involved in the CAP. As a result of their experience, ANS argues that all stakeholders involved should adjust to their educational philosophy and therefore also the CAP. As TT is still developing its CAP, it is very much aware of the aspects which should be improved before it can meet its quality standards, which are mentioned in their standards as

Table 3. Differences in reasoning about standards with examples from the group interview.

Criterion	Applied Natural Sciences	Teacher Training
Fitness for purpose	A CAP is fit for purpose when it is *aligned with the curriculum*. 'We've chosen to turn everything around, including the assessment. It is adjusted to the curriculum (...) so I'd say it is very much fit for purpose.'	A CAP is fit for purpose when there are *diverse assessment methods* that include *formative assessment*. 'Diverse assessment methods, yes ... that comes with competence-based education', '... formative assessment, which I think is conditional for good competence assessment'
Reproducibility of decisions	The responsibility for summative decisions should lie with the programme itself. 'We [educational institution] are still responsible.'	Hires professionals for summative assessments. 'The professional field is very happy to have more influence.'
Comparability	*Human deviations* from the specified assessment procedures are inevitable, but they do not influence the final summative decision. 'That will even out in the portfolio.'	The outcome of the specified assessment procedure for the final summative assessment should be *reliable*. 'We should use supervision ... to increase reliability.'
Fairness	A CAP is fair because deviations from the 'usual' content of the portfolio are possible. [if not the case] 'Then we would be back to the old system.'	A CAP is fair when students may also prove their competence with non-written proofs. 'Show us [the summative assessors] with other means that you are competent!'
Fitness for self-assessment	It is satisfactory when a CAP contains some instruments to practise self-assessment. 'I give it 45, while I consider it satisfactory. For the first year it is satisfactory.'	Students should shape their learning by selecting proofs in their portfolio. 'The students decide when and how to use formative assessments ... They need to develop their thinking about that.'
Cognitive complexity	The CAP content should require thinking steps. 'It is often the integrated assessments that fit best [with this criterion].'	The selection of proofs should also invite thinking steps. 'When all the contents of the portfolio are fixed ... you do not challenge students to think.'
Authenticity	Not all assessments have to be authentic. 'Let's not direct our efforts at that. Include a knowledge test. Use those means for something else.'	All assessment should be authentic. 'Then, should all assessments in the CAP be authentic?' 'I'd say so!'

conditions or prerequisites. As the TT stakeholders put it: 'if the formative assessment were organized properly' and 'with more training that should be possible'.

Secondly, professional orientation did not prove to be an explanation for systematic differences in the programmes' quality standards. References to the specific profession did not end up in the descriptive standards. This could be due to the fact that the programmes evaluated the CAPs of the first year of their educational programme. In discussing *Meaningfulness* and *Cognitive Complexity* the ANS participants in particular often commented that the profession was still far away. Therefore, it could be that professional orientation has an influence on quality standards for CAPs, but only in the final years of the educational programmes.

Thirdly, importance did not appear to have a consistent influence on the height of the cut-off scores or the descriptive standards. In general, during the group interview trade-offs between the different criteria were hardly discussed. However, there was one criterion whose standard clearly had an effect on the other standards, namely *Cost and Efficiency.* Although TT considered this criterion by far the least important of all the 12 quality criteria, and ANS regarded other criteria as more important as well, the criterion received priority in all the comparisons involving trade-offs. In that respect it appeared to function more as a bottleneck ('Can you ever consider Costs and Efficiency *too* much?') than as something which can increase CAP quality.

Besides these predetermined explanations, unexpected reasons for standard differences also emerged in the analysis, namely: (1) the quality of the actual CAP; (2) the use of self-evaluation as a method to specify standards; (3) the actual CAP's exemplary function; and (4) the understanding of the quality criteria. Firstly, ANS considered its actual CAP quality to be nearly satisfactory, whereas TT did not. This can be deduced by comparing the ratings for the current CAP to the specified numerical standards (see Table 2). This being the case, ANS could draw from concrete examples of CAP characteristics to specify its standards. TT only meets their numerical standards for *Authenticity* and *Educational Consequences*. In the group interview many aspects of their current CAP were deemed to be of unsatisfactory quality. This lack of concrete examples of satisfactory quality made it more difficult to specify a satisfactory CAP in a detailed way. This impression is supported by the fact that ANS was unable to specify a specific standard for the main aspect of its CAP that it considered unsatisfactory: its knowledge assessment. This issue came up many times during the interview and was often accompanied by comments such as: 'How should that be organized? That is difficult to describe', or 'I don't mean knowledge tests, but in some other way . . .'

Secondly, the self-evaluation procedure itself also had an influence on the resulting standards. Elements of the CAP that received considerable attention in the self-evaluation, often because they were deemed to be of unsatisfactory quality, also tended to appear in the standards. This was the case for ANS in the knowledge assessment example just mentioned. Another example is the TT's *Comparability* standard, which included reliable summative assessments. This was a prevalent topic of discussion among the TT staff and therefore had a large impact on the self-evaluation and the resulting standards.

Thirdly, and related to the first explanation, the actual CAP appeared to function as a powerful example for CAP standards. All four differences between the two CAPs resulted in standard differences (see Table 3 for examples). The appropriateness of two of these (i.e. the explicit exclusion of external assessors and of a definite list of contents for the portfolio in the ANS standards) was explicitly contemplated by ANS in the group interview, but the participants ruled in favour of their current CAP practices ('We've never done it that way, so why start now?').

Finally, the way in which the participants interpreted the criteria was also related to the differences in the standards. For instance, ANS focused on 'the ability to reflect, the conscious experience with what will come after [completion of] the programme' as part of *Authenticity*. TT did not include such elements but mainly appreciated *Authenticity* as the physical authenticity of its performance assessment, resulting in a different standard specification. Similarly, TT's standard specification of *Cognitive Complexity* included the complexity of the CAP itself – in which students had to select and shape their own proofs of competence – as a way to induce and measure thinking steps. ANS only included content complexity in their standard, in line with how the criterion was intended (see also Table 3).

Conclusion and discussion

The first goal of this study was to explore what level of compliance with CAP quality criteria two purposefully selected educational programmes consider satisfactory. This was done by explicating the standards implicitly applied in a CAP quality self-evaluation. This explication is expected to benefit the self-evaluation, as the actual situation can be compared to the standards, indicating possible improvements (Sadler 1998). A second goal of this study was to examine possible similarities and differences in the level and nature of the standards applied. The cross-case analysis showed that the two educational programmes specified similar cut-off scores, requiring a high level of compliance with all the quality criteria, but different descriptive standards, denoting different ways of how satisfactory quality should be operationalised. Our final aim was to establish possible influences on the level and nature of the standards. Based on our analyses, the standard differences resulted from the amount of experience with competence-based education and the quality and characteristics of the actual CAP. Also, the use of self-evaluation in the procedure and the understanding of the quality criteria were found to be of influence.

Some of these results warrant further discussion. These results indicate the existence of a general, to some extent inter-subjective understanding of what satisfactory quality for competence assessment entails (i.e. a similar level of compliance with the criteria), but a different perception of the manner in which such quality should be achieved (i.e. a different operationalisation). Based on this we reason that using numerical standards or cut-offs in standard setting gives an incomplete picture. These results also pinpoint the difficulties educational programmes are expected to experience in developing and subsequently evaluating their new assessments without available descriptive standards. This inability may have far-reaching consequences as assessment is a powerful determinant of student learning (Biggs 1996; Prodromou 1995). Hence, we would argue that introducing new types of assessment without available standards or examples, not uncommon in education, should be accompanied by some form of stakeholder self-evaluation.

Interestingly, the level of compliance with the criteria in the numerical standards contained little variation, as most scores were about .80. In contrast, the programmes did differentiate between the criteria in their descriptive standards. In these descriptive standards the level of compliance differed per criterion (i.e., not all ANS assessments have to be authentic, but all assessments should be comparable), but also between criteria. For instance, *Costs and Efficiency* received priority over the other criteria. Yet, although some of the specified standards conflict, consequential trade-offs were hardly discussed during the group interview. In the current study it remains unclear which standards would receive priority in considering trade-offs, as the measurement of importance assigned to the criteria did not capture that. As such trade-offs need to be considered when developing and implementing assessments (Haladyna and Hess 1999), this issue deserves more attention in further research.

There are limitations to our results. First of all, as these results are based on two cases, we believe the conclusions drawn can only be seen as a starting point for further research. Another point of critique may be that our procedure included both self-evaluation and standard explication. This triggered discussion about CAP elements of unsatisfactory quality. In turn, remnants of these discussions ended up in the standards, making them contingent on time and place and impeding their generalisation to other CAPs. This can be seen as an artefact of our study. Nevertheless, this effect has been described in the literature (Berk 1996; Hambleton et al. 2000; Price 2005; Sadler 1987) using different methods to specify standards.

The influence of the participant selection was also visible in our results. Their low agreement demonstrated the ANS stakeholders' diverging perspectives on satisfactory CAP quality, resulting from the different roles the participants have in the actual CAP. This effect is also described in the literature (McGinty 2005), but challenges conventional notions of appropriate agreement for standard setting, namely inter-rater reliability. On the other hand, as the stakeholders were selected for diversity, and assessment is currently seen to serve more than one purpose (Gielen, Dochy, and Dierick 2003; Linn, Baker, and Dunbar 1991) this result could also be appreciated. The reason why the TT stakeholders did not show these different perspectives in their cut-off scores may be that the ANS stakeholder group was more diverse. As a result of this and their more comprehensive perspective, the different perspectives on quality – resulting from different assessment purposes – were explicitly discussed in the ANS group interview. These results are an indication that an even more diverse stakeholder group, including, for instance, students and the professional field, will yield different results (Norcini and Shea 1997; Berk 1996). As these stakeholders are essential in quality assurance (Sadler 1987; Wolf et al. 1991; Vanhoof and Van Petegem 2007), especially in competence-based education, it may be wise to include them in future procedures to specify standards.

The understanding of the criteria also influenced the resulting standards. This effect has also been described before (Skorupski and Hambleton 2005; O'Donovan, Price, and Rust 2000). We incorporated the training session to impede it, but the current findings indicate that despite the training the understanding of the criteria differed. Hence, the multiple choice test measuring participants' understanding did not appear to be predictive of the way in which the participants applied the quality criteria further along the procedure. This could be the result of the overall complexity of the procedure – introducing stakeholders to a new level of analysis (i.e. that of a CAP instead of a single test) as well as 12 new quality criteria – but could also point to a more general issue. As Berk (1996) advocated in his review of standard-setting methods, in standard setting some frame of reference will always be used, which in turn may influence the resulting standards. As operationalisations of standards to fit the local situation are what make them meaningful for educational institutes, we postulate that the real challenge may not lie in reducing these purportedly erroneous effects. Yet, for quality assurance it is important to consider and subsequently account for local characteristics and understanding in the outcomes of standard-setting procedures.

This last issue runs through this study, and centres around the question of who determines the standards used in quality assurance. Our results indicate that when stakeholders are asked to set the standard, the level of the standards is similar across programmes. Yet, the descriptive standards appear to be grounded in local practices and therefore differentially operationalise how this assessment quality should be achieved. This adaptation to the local context makes the standards meaningful for the stakeholders, which in turn is likely to increase their ability to monitor and increase the assessment quality (Maslowski and Visscher 1999; Vanhoof and Van Petegem 2007). On the other hand, the same contextual nature of the standards might induce problems for the comparability between programmes, which is often the goal of external evaluations (Nevo 2001).

Paradoxically, in the current educational landscape there appear to be trends towards self-evaluation, including using internal quality expectations *and* accountability with external standards at the same time (McNamara and O'Hara 2008). As the advantages of both, ownership and comparability, are appealing, it may be worthwhile to combine them (Nevo 1994). Based on our results, we argue that the shared reflection by stakeholders on what constitutes satisfactory quality should be seen as a powerful tool in development of innovative assessments and, perhaps, as part of quality assurance in itself. However, to guarantee that different operationalisations of the level of compliance with the quality criteria in fact constitute the same quality, we agree with Vanhoof and Van Petegem (2007) that some degree of external support and monitoring should be in place. On the one hand, this support should focus on awareness of what constitutes quality and how it can be achieved, for instance by providing powerful exemplars, as argued by McNamara and O'Hara (2008). On the other hand, we feel the real challenge may lie in establishing a system of monitoring that explores new ways of assuring the quality of internal evaluations and standard-setting procedures. This monitoring system could focus on the strength of the evidence and arguments for the assessment quality provided by stakeholders, as an alternative to inspecting concrete assessment practices.

Notes on contributors

Larike H. Bronkhorst (MSc) is a junior researcher at the Department of Education, Faculty of Social and Behavioural Sciences, Utrecht University. Her research focuses on student teacher learning, deliberate practice and assessment of learning.

Liesbeth K.J. Baartman (PhD) is a post-doctoral researcher at the Eindhoven School of Education at Eindhoven University of Technology. She specialises in research on assessment, quality criteria for assessment and competence-based (vocational) education. Currently, she is involved in a research project into what competences employees of the twenty-first century need, in the domain of science and technology specifically.

Karel M. Stokking is professor of Education at the Department of Education, Faculty of Social and Behavioural Sciences, Utrecht University. His areas of interest include the development and assessment of general and academic skills in secondary and higher education.

References

Akkerman, S., W. Admiraal, M. Brekelmans, and H. Oost. 2008. Auditing quality of research in social sciences. *Quality and Quantity* 42, no. 2: 257–74.

Allen, D., and K. Tanner. 2006. Rubrics: Tools for making learning goals and evaluation criteria explicit for both teachers and learners. *CBE Life Science Education* 5, no. 3: 197–203.

Baartman, L.K.J., T.J. Bastiaens, P.A. Kirschner, and C.P.M. Van der Vleuten. 2006. The wheel of competency assessment: Presenting quality criteria for competency assessment programs. *Studies In Educational Evaluation* 32, no. 2: 153–70.

Baartman, L.K.J., F.J. Prins, P.A. Kirschner, and C.P.M. Van der Vleuten. 2007. Determining the quality of competence assessment programs: A self-evaluation procedure. *Studies in Educational Evaluation* 33, nos. 3–4: 258–81.

Baartman, L.K.J., F.J. Prins, P.A. Kirschner, and C.P.M. van der Vleuten. 2011. Self-evaluation of assessment programs: A cross-case analysis. *Evaluation and Program Planning* 34, no. 3: 206–16.

Bennett, Y. 1993. The validity and reliability of assessments and self-assessments of work-based learning. *Assessment & Evaluation in Higher Education* 18, no. 2: 83–95.

Berk, R.A. 1996. Standard setting: the next generation (where few psychometricians have gone before!). *Applied Measurement in Education* 9, no. 3: 215–35.

Biggs, John. 1996. Enhancing teaching through constructive alignment. *Higher Education* 32, no. 3: 347–64.

Birenbaum, M. 1996. Assessment 2000: Towards a pluralistic approach to assessment. In *Alternatives in assessment of achievement, learning processes and prior knowledge*, ed. M. Birenbaum and F.J.R.C. Dochy, 3–29. Boston, MA: Kluwer Academic.

Birenbaum, M. 2003. New insights into learning and teaching and their implications for assessment. In *Optimising new modes of assessment: In search of quality and standards*, ed. M. Segers, F. Dochy, and E. Cascallar, 13–36. Dordrecht, The Netherlands: Kluwer Academic.

Birenbaum, M. 2007. Evaluating the assessment: Sources of evidence for quality assurance. *Studies in Educational Evaluation* 33, no. 1: 29–49.

Delandshere, Ginette. 2001. Implicit theories, unexamined assumptions and the status quo of educational assessment. *Assessment in Education: Principles, Policy & Practice* 8, no. 2: 113–33.

Dochy, Filip. 2001. A new assessment era: Different needs, new challenges. *Learning and Instruction* 10, no. S1: 11–20.

Dochy, F.J.R.C., and Liz McDowell. 1997. Assessment as a tool for learning. *Studies in Educational Evaluation* 23, no. 4: 279–98.

Downing, Steven M. 2003. Validity: On the meaningful interpretation of assessment data. *Medical Education* 37, no. 9: 830–7.

Driessen, E., C. Van der Vleuten, L. Schuwirth, J. Van Tartwijk, and J. Vermunt. 2005. The use of qualitative research criteria for portfolio assessment as an alternative to reliability evaluation: A case study. *Medical Education* 39, no. 2: 214–20.

Erkens, G. 2005. *Multiple Episode Protocol Analysis (MEPA)*. Version 4.10. Utrecht, The Netherlands: Utrecht University.

Gielen, S., F. Dochy, and S. Dierick. 2003. Evaluating the consequential validity of new modes of assessment: The influence of assessment on learning, including pre-, post-, and true assessment effects. In *Optimizing new modes of assessment: In search of qualities and standards*, ed. M. Segers, F. Dochy, and E. Cascallar, 37–54. Dordrecht, The Netherlands: Kluwer Academic.

Greene, David, and Jane L. David. 1984. A research design for generalizing from multiple case studies. *Evaluation and Program Planning* 7, no. 1: 73–85.

Haladyna, T., and R. Hess. 1999. An evaluation of conjunctive and compensatory standard-setting strategies for test decisions. *Educational Assessment* 6, no. 2: 129–53.

Hambleton, R.K., R.M. Jaeger, B.S. Plake, and C.N. Mills. 2000. Setting performance standards on complex educational assessments. *Applied Psychological Measurement* 24, no. 4: 355–66.

Kane, M. 1994. Validating the performance standards associated with passing scores. *Review of Educational Research* 64, no. 3: 425–61.

Knight, P. 2000. The value of a programme-wide approach to assessment. *Assessment & Evaluation in Higher Education* 25, no. 3: 237–51.

Koretz, D., and E. Deibert. 1995. Setting standards and interpreting achievement: A cautionary tale from the National Assessment of Educational Progress. *Educational Assessment* 3, no. 1: 53–81.

Lichtman, M. 2006. *Qualitative research in education. A user's guide*. Thousand Oaks, CA/London/New Delhi: Sage.

Linn, R.L., E.L. Baker, and S.B. Dunbar. 1991. Complex, performance based assessment: Expectations and validation criteria. *Educational Researcher* 20, no. 8: 15–21.

Lizzio, A., and K. Wilson. 2004. Action learning in higher education: An investigation of its potential to develop professional capability. *Studies in Higher Education* 29, no. 4: 469–88.

Maslowski, Ralf, and Adrie J. Visscher. 1999. Formative evaluation in educational computing research and development. *Journal of Research on Computing in Education* 32, no. 2: 239–55.

McGinty, D. 2005. Illuminating the "Black Box" of standard setting: An exploratory qualitative study. *Applied Measurement in Education* 18, no. 3: 269–87.

McNamara, G., and J. O'Hara. 2008. The importance of the concept of self-evaluation in the changing landscape of education policy. *Studies in Educational Evaluation* 34, no. 3: 173–9.

Nevo, David. 1994. Combining internal and external evaluation: A case for school-based evaluation. *Studies in Educational Evaluation* 20, no. 1: 87–98.

Nevo, David. 2001. School evaluation: Internal or external? *Studies in Educational Evaluation* 27, no. 2: 95–106.

Norcini, J.J., and J.A. Shea. 1997. The credibility and comparability of standards. *Review of Research in Education* 17: 3–29.

O'Donovan, B., M. Price, and C. Rust. 2000. The student experience of criterion-referenced assessment (through the introduction of the common criteria assessment grid). *Innovation in Education and Teaching International* 38, no. 1: 74–85.

O'Donovan, B., M. Price, and C. Rust. 2004. Know what I mean? Enhancing student understanding of assessment standards and criteria. *Teaching in Higher Education* 9, no. 3: 325–35.

Price, M. 2005. Assessment standards: The role of communities of practice and the scholarship of assessment. *Assessment & Evaluation in Higher Education* 30, no. 3: 215–30.

Prodromou, L. 1995. The backwash effect: From testing to teaching. *ELT Journal* 49, no. 1: 13–25.

Purves, A.C. 1993. Setting standards in the language arts and literature classroom and the implications for portfolio assessment. *Educational Assessment* 1, no. 3: 174–99.

Sadler, D.R. 1987. Specifying and promulgating achievement standards. *Oxford Review of Education* 13, no. 2: 191–209.

Sadler, D.R. 1989. Formative assessment and the design of instructional systems. *Instructional Science* 18, no. 2: 119–44.

Sadler, D.R. 1998. Formative assessment: Revisiting the territory. *Assessment in Education: Principles, Policy & Practice* 5, no. 1: 77–85.

Scriven, M. 1980. *The logic of evaluation*. Iremess, CA: Edgepress.

Skorupski, W.P., and R.K. Hambleton. 2005. What are panelists thinking when they participate in standard setting studies? *Applied Measurement in Education* 18, no. 3: 233–56.

Tillema, H.H., J.W.M. Kessels, and F. Meijers. 2000. Competencies as building blocks for integrating assessment with instruction in vocational education: A case from The Netherlands. *Assessment & Evaluation in Higher Education* 25, no. 3: 265–78.

Tynjälä, Päivi. 1999. Towards expert knowledge? A comparison between a constructivist and a traditional learning environment in the university. *International Journal of Educational Research* 31, no. 5: 357–442.

Van Berkel, H.J.M., and H.A.P. Wolfhagen. 2002. The Dutch system of external quality assessment: Description and experiences. *Education for Health* 15, no. 3: 335–45.

Van Der Schaaf, Marieke F., Karel M. Stokking, and Nico Verloop. 2003. Developing performance standards for teacher assessment by policy capturing. *Assessment & Evaluation in Higher Education* 28, no. 4: 395–410.

Van der Vleuten, C.P.M., and L.W.T. Schuwirth. 2005. Assessing professional competence: From methods to programmes. *Medical Education* 39, no. 3: 309–17.

Vanhoof, J., and P. Van Petegem. 2007. Matching internal and external evaluation in an era of accountability and school development: Lessons from a Flemish perspective. *Studies in Educational Evaluation* 33, no. 2: 101–19.

Wolf, D., J. Bixby, J. Glenn, and H. Gardner. 1991. To use their minds well: Investigating new forms of student assessment. In *Review of research in education*, ed. G. Grant, 31–74. Washington, DC: American Educational Research Association.

Appendix 1. Descriptive quality standards per programme

Criterion	Programme	Verbal description of quality standard
Fitness for purpose	ANS	All forms of assessment are aligned with competence-based education and the educational philosophy. All competences are assessed, with sufficient attention for knowledge, skills, and attitude as well as their integration. Separate knowledge assessments also take place, not just with a knowledge test, but also in other ways to see if students possess the necessary knowledge. The CAP prepares the students to compensate for a potential knowledge gap, now and in the future.
	TT	The assessment tasks contain variation. The CAP contains formative and summative assessment. The TT assessment matrix specifies which proofs of competence the portfolio needs to have, without specifying a specific TT domain. Knowledge assessment weights high, by means of a separate test or as part of an integrated assessment. When integrated competences can be assessed, no separate knowledge tests are necessary. Skills and attitudes are assessed sufficiently.
Reproducibility of decisions	ANS	Summative decisions are based on several proofs of competence, gathered from different situations and different assessors. These assessors have different backgrounds. Professional assessments are important, but internal assessors should make the summative decisions. Summative assessors need not discuss their opinions about the portfolio with the formative assessors if the formative assessors follow the specified assessment procedure. Deviations in formative assessments will average out in the students' portfolio.
	TT	Different proofs of competence gathered under different conditions assessed by different assessors are required to prove competences. Formative and summative assessors do not discuss their opinions; the latter base their final decision on the comments of the former, without reassessing the formative proofs. Summative assessors prepare the portfolio assessment separately, but discuss their final decision.
Transparency	ANS	The CAP is specified on paper, but there is not a fixed list as to what proofs should be in the portfolio to give students more freedom. Every year students need some time to understand the CAP and ANS supports them in this process with instruction and practice assessments. Teachers know and understand the CAP. The professional field understands its main points.
	TT	The CAP is completely specified on paper. Students understand the CAP, if necessary through extra guidance and support. Teachers understand the CAP. The professional field need not understand the CAP completely to execute their part of the assessments, but will realise a more profound understanding with more experience.

(*continued*)

Appendix 1. (*Continued.*)

Criterion	Programme	Verbal description of quality standard
Acceptance	ANS	Students need and are granted time to adjust to and accept the educational philosophy and therefore the CAP every year and are supported in this process by means of instruction and practice assessments. Teachers accept the CAP. The professional field was involved in establishing the end goals of the ANS programme to ensure that they accept it. All stakeholders have confidence in the CAP.
	TT	Students, teachers and the professional field accept the CAP, as they have considerable experience with it, the formative assessment is of high quality and the different TT domains have explicit roles in it. Stakeholders have confidence in the CAP, because it secures the knowledge bases of students and guarantees the quality of the summative assessors.
Comparability	ANS	All assessments are comparable as they take place under comparable circumstances and are assessed on equal criteria and standards in a fixed assessment procedure. The content of the student portfolio is not specified, nor fixed. Assessors are human, so there can be small differences in their assessments, which average out in the portfolio and do not influence the final/summative decision.
	TT	The fixed parts of the TT assessment matrix are comparable; the proofs of competence students select themselves need not be. The instruction towards the competence assessment is based on the same criteria and standards and a prescribed assessment procedure. The content of the student portfolio is specified. The final summative decision is reliable.
Fairness	ANS	Students can appeal against assessments. If they have complaints or remarks, these are heard and possibly met. The CAP offers students sufficient possibilities to prove competence, as there is no fixed list for the portfolio. The CAP is comparable, making a fair summative decision possible. Summative assessors are trained and formative assessors are either trained or supported in their assessment task.
	TT	Students can appeal against assessments. Complaints are used to improve the CAP. Students consider the CAP to be fair, as it offers them sufficient possibilities to prove their competence, without relying on writing abilities. Formative assessors are trained and their work is subjected to supervision. Summative assessors are certified. They assess in different duos.

Criterion	Programme	Verbal description of quality standard
Fitness for self-assessment	ANS	The CAP is based on self-assessment. The CAP contains various instruments with which students learn to shape their own learning. They practise self- and peer assessment. Reflection on competence development in the portfolio is necessary to prove proficiency. Students formulate their own learning goals, triggered by practice assessments, among other things. Teachers give good feedback on a regular basis.
	TT	Students can shape their own learning in the curriculum and the CAP itself, as the TT assessment matrix requires proofs of competence students select themselves. This invites them to reflect and thereby shape their own learning. To avoid too much complexity, students are supported in this process. Supported by assessors, students learn to reflect, especially in the criterion-based interview. Students also practise with self- and peer assessment in giving and receiving feedback. Assessors provide students with high quality feedback on a regular basis.
Meaningfulness	ANS	Students consider the assessments and accompanying feedback meaningful for their learning process, but do not make a direct link to their future profession. They see assessments as 'learning moments' and use the feedback given. Teachers and the professional field consider the CAP to be useful for both learning process and future occupation. The professional field was involved in establishing the end goals of the curriculum to ensure its meaningfulness.
	TT	Students see the feedback of assessments as meaningful for their learning process and consider both formative and summative assessments as 'learning moments', steering their learning process in the right direction. Teachers and the professional field consider the CAP to be meaningful for students' learning process and future profession.
Cognitive complexity	ANS	The curriculum and the CAP are cognitively complex, as students have to make a plan of action and intermediate reports, to learn how to communicate within the professional field and to analyse complex problems. Thinking steps receive explicit attention in several assessments, although the summative assessment criteria focus on competences. The thinking level of the CAP is derived from that required for a professional, but adjusted to the first year of the programme. The application of knowledge is also satisfactorily assessed.
	TT	Students need to select and shape their own proofs of competence and learning process in the TT assessment matrix, which requires thinking steps. Students explain these steps at assessments, especially at the criterion-based interview, where they link theory to practice. Assessment criteria specify knowledge development, next to behaviour. The required thinking level of the CAP may differ in the first year, as students may require extra meta-cognitive support, but the end level of the TT programme is equal to that of a newly starting professional.

(*continued*)

Appendix 1. (*Continued.*)

Criterion	Programme	Verbal description of quality standard
Authenticity	ANS	The assessments and assessment criteria resemble those of the future profession, but adjusted to the first year of the programme. Not all assessments are authentic if this conflicts with *Costs and Efficiency*. The physical circumstances of the performance assessment are authentic. In other assessments authentic elements such as co-operating, contact with a superior, and refusing are present.
	TT	All assessments are authentic. The physical and social circumstances of the CAP and the elements of the instruction resemble those of the future profession, for instance, when the students reflect on practice. Assessment criteria specify concrete actions of a professional, are established in consultation with the professional field and are used there as well.
Educational consequences	ANS	Practice assessments confront students with their progress. Together with other assessments and instruction, they motivate the students more than the final/summative assessment, which takes place too late to adequately do that. The CAP has little influence on instruction, although the results of the CAP are used to adjust the curriculum in combination with other evaluations.
	TT	Indicators of assessment criteria specify behaviour and knowledge development. The students are motivated by good feedback. Formative assessments are valued, as students use them to measure their progress. The CAP influences the methods of instruction, not the content. Consistent adjustments are made to the whole programme CAP based on the results of the CAP, as well as other evaluations.
Costs and efficiency	ANS	An estimate of the costs and efficiency of the CAP is made prior to its implementation and forms the basis for the designation of the mix of assessment methods. Some assessments are consciously omitted or are not offered too frequently. The CAP's efficiency is regularly evaluated to constantly increase it. Students can complete the CAP reasonably. Teachers have enough time for assessments and feedback.
	TT	The mix of assessment methods is determined by an estimation of the costs and efficiency. It is evaluated whether the CAP is efficient enough, and if its costs outweigh the learning gain. Teacher training and student guidance make the CAP reasonable to execute.