Running head: QUALITY CRITERIA

**This article was published as**

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



# Teachers' opinions on quality criteria for Competency Assessment Programmes

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

[a] Open University of the Netherlands

[b] Maastricht University, the Netherlands

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

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

Abstract

Quality control policies towards Dutch vocational schools have changed dramatically because the government questioned examination quality. Schools must now demonstrate assessment quality to a new Examination Quality Center. Since teachers often design assessments, they must be involved in quality issues. This study therefore explores teachers' opinions on assessment quality evaluation criteria. Pre-vocational and vocational teachers (N = 211) responded to a questionnaire. Contrary to expectations, results show that teachers deem classical and competency-based quality criteria equally important. Vocational teachers gave higher importance scores than pre-vocational teachers, possibly due to the pressure they experience to improve the quality of their assessments.

Keywords: evaluation criteria, alternative assessment, teacher attitudes, vocational education

# Teachers' Opinions on Quality Criteria for Competency Assessment Programs

There is a strong pressure on educational institutes and teachers to become more competency-based, so as to better meet the changing demands of the labour market. This has important consequences for student assessment because of the strong relationship that exists between instruction, learning and assessment. Assessment, learning and instruction should be aligned with each other (i.e. focus on the same learning outcomes). Also, assessment appears to strongly influence both how students learn and how teachers teach, causing both students and teachers to focus on what the assessment requires (e.g. Alderson & Wall, 1993; Biggs, 1999; Birenbaum, 2003; Frederiksen, 1984). A study focusing on how teachers connect instruction and assessment showed that teachers spend more than 35 % of their time on assessment and more than 10 % on assessment-driven instruction (Conca, Schechter, & Castle, 2004). A possible problem here is that whereas learning and instruction are increasingly competency-based, the development of adequate methods to assess those competencies appears to be lagging behind. The past decade has seen a number of new assessment forms, such as performance assessment, authentic assessment and portfolio assessment (Gulikers, Bastiaens, & Kirschner, 2004; Hambleton, 1996; McDowell, 1995), each of which promises a panacea for the assessment of competencies. But because competencies are so difficult to assess, using one single assessment form seems not to be sufficient (Chester, 2003). Based on earlier work of the authors (Baartman, Bastiaens, Kirschner, & Van der Vleuten, in press), this article argues for a combination of different assessment forms - a Competency Assessment Program (CAP)[1] - which combines both classical tests and recently developed assessment methods.

The use of CAPs in competency-based education seems promising, but teachers and educational institutes are struggling with how to determine the quality of the different assessment forms they use, both individually and in combination. Many teachers believe that they need strong measurement skills to construct assessments, and report a level of discomfort with the quality of their own assessments (Frey, Petersen, Edwards, Pedrotti, & Peyton, 2005). Two

reasons can be given for this struggle. First, criteria such as validity and reliability, which have long been sufficient for classical testing are necessary, may not be sufficient for new assessment forms and combinations of these forms in CAPs (Moss, 1994; Taylor, 1994). Moreover, validity and reliability are defined and used in many different ways (Miller & Linn, 2000), which makes it difficult for teachers to effectively implement them in practice to evaluate their assessment methods. Maclellan (2004) showed that among novice teachers there was very little exemplification or elaboration of the concepts of validity and reliability and they did not connect issues of reliability and validity with different assessment methods. With the development of new assessment forms, concomitant, complementary or supplementary quality criteria have been proposed, such as the consequences, meaningfulness and cognitive complexity of the assessment (e.g., Linn, Baker, & Dunbar, 1991; Kane, 1992, 2004; Van der Vleuten & Schuwirth, 2005). Since CAPs consist of combinations of both classical tests and newly developed assessment forms, quality criteria from both classical and new views on quality might be needed to evaluate their effectiveness.

Second, apart from the fact that from a theoretical point of view it is unclear what quality criteria should be used for CAPs, educational institutes in the Netherlands are increasingly held responsible for demonstrating the quality of their assessments and are, therefore, looking for adequate criteria to evaluate their CAPs. An interesting case in this respect is vocational education. In the Netherlands, after leaving primary school, all pupils are required to enter secondary education. Here, they choose between general secondary education, which leads to entrance to a university of polytechnic, and pre-vocational education (age 12-15). Pre-vocational education servers as preparation for vocational education, which can be taken at a range of levels (age 15-18). The Dutch vocational schools are comparable to the American vocational high schools. Almost half of the yearly cohort of Dutch pupils leaving primary school eventually enters some form of vocational education. When finishing vocational education, pupils choose either to enter higher professional education – comparable to vocational colleges or polytechnics

– or to receive a vocational certification. In 2001, the Dutch government expressed little trust in the quality of the examinations in schools for vocational education (Deetman, Stuurgroep Examens, 2001). To improve quality, the Examination Quality Center was established in 2004, which defined national standards for quality to which vocational schools must conform in order to retain their accreditation. In this vision, it is the schools who are responsible for demonstrating that their examinations meet those standards[2]. If the standards are met, the school receives its accreditation from the Examination Quality Center, which allows to examine and certify their students. Without such accreditation, schools must enlist the services of another accredited institution. On top of this requirement, external monitoring has been increased to cover 100% of all examinations. The quality standards used by the Examination Quality Center focus on: management and organization of examination, contracting out examinations, examination process, examination products and accountability.

There is a problematic dichotomy here. First, the onus of proof of quality is shifted to schools, but schools seem not to be well-equipped – as an institution – to carry out this quality control. Vocational schools are struggling between the strict and often classical standards set by the Examination Quality Center and their wish to make education more competency-based (Onderwijsraad, 2006). Because teachers in vocational schools often design assessments, responsibility for quality control is also passed on to them. Second, since the teacher too is not especially qualified to carry out quality control, their individual credibility is threatened. The evaluation of assessment programs is usually carried out by school management and external controlling bodies without involving the teachers working at the schools, although they are an important factor for achieving high quality CAPs. In the Netherlands, it is the teacher who actually develops and carries out assessments and who has to make sure quality is well established. On the other hand, teachers have to work within an area of accountability and external control, which may threaten their credibility as teachers capable of their own assessment of student learning (Graham, 2005).

~~2-2-2009~~

To assist both schools and their teachers, useful and usable quality criteria for assessments are needed. The goal of this study is to explore teachers' opinions on quality criteria for CAPs. In a previous study, a framework of ten quality criteria for CAPs was developed, which was validated by means of an expert focus group meeting (Baartman et al., in press). This framework is shortly described in the next section. The current study focuses on the validation of this framework by the actual users and developers of the assessment programs, the teachers.

*Ten-Criterion Framework for CAPs*

Our framework of quality criteria is based on a synthesis of work by many authors (e.g. Driessen, Van der Vleuten, Tartwijk, & Vermunt, 2005; Gulikers, et al., 2004; Hambleton, 1996; Linn et al., 1991; Kane, 2004; Schuwirth & Van der Vleuten, 2004; Uhlenbeck, 2002; Van der Vleuten & Schuwirth, 2005). The goal of the framework is to provide a clear definition of all relevant criteria to enable their further operationalisation into an instrument for schools and teachers for evaluating CAPs. The framework comprises both criteria related to the classical ideas of quality control such as comparability, fairness, reproducibility of decisions and transparency, and criteria that arose during the transition towards competency-based education such as authenticity, cognitive complexity, costs and efficiency, directness and educational consequences, meaningfulness. Since CAPs consist of combinations of assessment methods, it is important to note that not all single methods included in a CAP must meet all criteria, but that the CAP as a whole must. For example, a non-authentic assessment form such as a written test for assessing knowledge about nurse-patient communication can be combined with a more authentic assessment form such as a performance assessment, in which the student really has to show his or her capabilities in communicating with patients. A CAP as a whole, on the other hand, has to comply with all quality criteria. For example, high scores on authenticity cannot offset deficits in cognitive complexity. Table 1 gives a short description of the ten criteria. For a more elaborate description see Baartman et al. (in press).

**- INSERT TABLE 1 ABOUT HERE -**

~~2-2-2009~~

This framework of quality criteria has already been validated by experts in the field of assessment and quality criteria for assessment (Baartman et al., in press). The goal of this study is to explore the opinion of a second important group of stakeholders in the assessment process, the teachers. First, the study investigates whether teachers consider quality criteria to be important for evaluating their assessment programs and second, whether they deem some criteria more important than others. We expected teachers to deem classical quality criteria more important than newer competency-based criteria, as teachers are often thought to be reluctant towards this change to competency-based education and assessment. The distance between school managers and teachers seems to be increasing, causing teachers to only focus on their primary task of teaching, resulting in less commitment and awareness towards educational change (Onderwijsraad, 2006). The third goal of this study is to compare the views of teachers working in different types of education with different quality control policies. As described in the introduction, quality control policies in vocational education have changed dramatically in the Netherlands in the last half decade. This study compares the views of teachers working in vocational education to those of teachers working in pre-vocational education, the type of education leading towards vocational education. In 2001, a group of technical pre-vocational schools (called the ICT-route) got permission from the Dutch Ministry of Education to develop a new curriculum and assessment for technical pre-vocational education. They developed their own assessment program, focusing on formative assessment and, working together with vocational schools, strived to permit a more fluid transition between pre-vocational and vocational education (Van der Sanden, Van Os, & Kok, 2003). Because assessment in these pre-vocational schools has a more formative and competency-based character, we expected teachers from these schools to deem newer quality criteria more important, whereas we expected teachers from vocational schools to deem classical criteria more important. Using a questionnaire, teachers' opinions about the quality criteria were investigated and differences between vocational and pre-vocational education were studied.

~~2-2-2009~~

## Method

### *Participants*

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

### *Materials*

A questionnaire was developed based on the ten quality criteria of the theoretical study, in which the teachers were asked about the importance of the ten quality criteria for their assessments. The questions covered the theoretical definitions and descriptions of the quality criteria. As the quality criteria are fairly abstract concepts, the questions were formulated as examples of the quality criteria in practice. For example, in one of the *authenticity*-questions, teachers were asked whether they deem it important to assess students in the workplace. Scales of four to eight questions were composed for each quality criterion. For the criterion *cognitive complexity* two subscales were developed, namely thinking processes and thinking level. The criterion Thinking processes deals with the assessment of the way students think, make decisions, and provide a rationale for their decisions when performing a task. Thinking level pertains the difficulty of the cognitive skills needed to solve the problems encountered on the job. At the end of the questionnaire, an open question enabled the teachers to give further comments on the quality of assessments and their experiences with it. Table 2 presents the scales, the number of items in each scale and an example of an item of each scale. Answers on

all questions were given on a 5-point Likert scale ranging from (1) not important at all to (5) very important. The last question was an open one in which the respondents were asked to freely express their opinion on quality criteria for assessments. In the instruction accompanying the questionnaire an explanation was given of a Competency Assessment Program and the teachers were encouraged to give their personal opinion about the importance of the criteria: "please give your personal opinion as a teacher, independent of current assessment practices and policy at your school. We would like to know what competency assessment should look like in your opinion".

**- INSERT TABLE 2 ABOUT HERE -**

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

Before analyzing the results of the questionnaire, the reliability of the criterion scales was determined. Table 2 also shows the Cronbach's Alpha scores of all scales, which were found to be moderately to highly reliable (range .59 to .82). To increase scale reliability, one question with a low item-total correlation value was removed from the *transparency* scale. To explore whether the criterion scales were uni-dimensional, a factor analysis was conducted on each scale. All scales proved to be uni-dimensional except for *cognitive complexity*. As was expected, this scale was composed of two distinct subscales. A factor analysis with Varimax rotation showed two factors, with Eigenvalue 3.57 and loading ranging from .487 to .866 for *thinking level* and Eigenvalue 1.05 and loading ranging from .661 to .737 for *thinking processes*. The first factor consists of all questions regarding thinking level and the second factor includes the questions about assessing thinking processes.

*Procedure*

~~2-2-2009~~

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

*Analysis*

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

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

## Results

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

*Perceived Importance of the Quality Criteria*

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

vocational education (M ranging from 3.68 to 4.35, $p < .001$ for all criteria) and vocational education (M ranging from 3.96 to 4.54, $p < .001$) separately.

**- INSERT TABLE 3 ABOUT HERE -**

The ANOVA yielded a significant main effect (Greenhouse-Geisser, $F$ (6.92, 1439.83) = 12.89, $MSE$ = .38, $\eta_p^2$ = .058, $p < .001$), indicating differences in importance scores between the criteria. Post hoc tests (Bonferroni) were used to further investigate the differences between the criteria. Figure 1 shows the mean importance scores given by the teachers, together with the 95% confidence interval of the comparison between the different criteria. For easier comparison, the criteria in the figure have been ordered from most to least important.

**- INSERT FIGURE 1 ABOUT HERE –**

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

> My personal opinion is that each student enters the school with a number of competencies. Our goal is to stimulate the student to develop these competencies en to teach competencies the student is less interested in. We have to assess what the student learns, and not what has to be learned. The personal interests of each student should guide the assessment. Each student should be able to get a certificate/diploma based on his or her competencies. Assessment thus has to be very personalized.
>
> ~~2-2-2009~~

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

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

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

*Differences between educational levels*

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

**- INSERT FIGURE 2 ABOUT HERE -**

~~2-2-2009~~

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

## Conclusion and Discussion

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

The first research question focused on whether teachers considered the quality criteria to be important. The results show that this is indeed the case. As expected, all quality criteria were given high to very high scores on the importance scale, showing that teachers think it is important they use high-quality CAPs. On the other hand, this does not mean they also actually carry out quality checks. As did the experts in our previous study (Baartman et al., in press), the users of CAPs apparently consider the quality criteria to be relevant, hereby validating the framework.

With regard to the second research question, the results show that teachers consider classical criteria and newer criteria to be equally important. This is interesting, as teachers are often thought to be reluctant towards adopting new assessment methods and criteria (Onderwijsraad, 2006). The discussion about whether or not it is necessary to complement the classical views on quality control with new quality criteria has been going on for some time within the scientific field (e.g., Bachman, 2002; Moss, 1996; Webb, Endacott, et al., 2003). This

study adds a different point of view, that of teachers, to this discussion. The results seem to support the idea of combining both classical and new views on quality control into an integral quality framework for CAPs. Also, the results show that, while all criteria were considered important, some criteria were deemed more important than others. *Transparency* scored very high, which may be due to the fact that in vocational education it was stressed by the Examination Quality Center during their audits in the preceding years. The government's critics (Deetman, Stuurgroep Examens, 2001) on the vocational examinations also addressed the lack of transparency and comparability between institutes. One of the main tasks of this new Examination Quality Center was to improve transparency. Whereas until 2001 they only evaluated 50 % of all summative examinations carried out at a school, they now check all of them, hereby expressing a clear wish to gain better insight in the assessment practices carried out at vocational schools. Increased transparency was needed for them to be able to achieve this insight. A second explanation of the high scores on *transparency*, which could apply to pre-vocational education, is that, being in a transition period towards competency-based education, teachers experience many uncertainties in their work as a teacher, which increases their need for clarity about assessments. *Reproducibility* on the other hand scored relatively low compared to the other criteria. Assessing each student in different situations and the use of multiple assessors is often considered to be a possible solution to the reliability problems faced in competency-based education, but apparently teachers thought this relatively less important than other quality aspects. An explanation for the lower scores on *reproducibility* could be that teachers are not used to assessing students together with colleagues or other people and are afraid of loosing their autonomous position. Being professional teachers, they possibly regard themselves as objective judges, hereby mistaking being a professional for automatically being objective. Another possibility is that the use of multiple assessors is just not a habit in vocational schools or teachers might think it is not feasible, being too costly and time-consuming.

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

Teachers in vocational education gave higher importance scores on *costs & efficiency*, *cognitive complexity* – thinking level and *fairness*. The higher scores on *costs & efficiency* indicate that teachers in schools are more concerned that new assessment methods will be too expensive and too time-consuming. Until recently, these schools have had less opportunity to experiment with new assessment methods, which might explain this reluctance. The results also indicate that giving schools the opportunity and freedom to experiment with new assessment methods, as was done in our group of pre-vocational schools, could diminish reluctance towards these innovations. The fact that vocational education teachers judge *cognitive complexity* – thinking level to be more important can be explained by the fact that they are working at a higher level of education. As stated, pre-vocational education is not the end station for most students, while at the end of vocational education most students start working. At the end of pre-vocational education, the thinking level is still less important than it is at the end of vocational

education, because this is when students need to be prepared to start working in a specific professional field with the accordingly required level of reasoning, or continue their education in institutions for higher vocation education, which also poses demands on thinking level. The higher *fairness* scores given by teachers in vocational education is to be expected since assessment in vocational education is generally meant for certification, whereas assessment in pre-vocational education is not. In the eyes of teachers, *fairness* is probably more important in summative than in formative assessment situations. Further research is needed here into the question whether the same quality criteria should apply for formative and summative assessments.

To conclude, this study presents teachers opinions on a framework of quality criteria for CAPS, which includes both classical and new views on assessment. As such,it adds a different point of view, that of teachers, to the scientific discussion about quality criteria for competency assessment. The framework provides an answer to the discussion about whether or not it is necessary to complement the classical views on quality control with new quality criteria that may do more justice to the unique character of competency assessment. In this research, both views are combined into an integral framework and, just like experts (Baartman et al., in press), teachers appear to support this idea.

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

~~2-2-2009~~

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

References

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

Baartman, L. K. J., Bastiaens, T. J., & Kirschner, P. A. (in press). The wheel of competency assessment. Presenting quality criteria for Competency Assessment Programmes. *Studies in Educational Evaluation, 32*.

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

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

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

Birenbaum, M. (2003). New insights into learning and teaching and their implications for assessment. In M. Segers, F. J. R. C. Dochy, & E. Cascallar (Eds.), *Optimising new modes of assessment: In search of qualities and standards* (pp. 13-36). Dordrecht, The Netherlands: Kluwer Academic Publishers.

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

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

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

~~2-2-2009~~

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

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

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

Gipps, C. V. (1994). *Beyond testing: Towards a theory of educational assessment.* London: The Falmer Press (Taylor & Francis Group).

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

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

Hambleton, R. K. (1996). Advances in assessment models, methods, and practices. In D. C. Berliner & R. C. Calfee (Eds.), *Handbook of Educational Psychology* (pp. 899-925). New York: MacMillan.

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

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

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

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

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

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

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

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

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

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

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

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

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

Uhlenbeck, A. M. (2002). *The development of an assessment procedures for beginning teachers of English as a foreign language.* Unpublished doctoral dissertation, University of Leiden, ICLON Graduate School of Education, Leiden, The Netherlands.

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

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

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

Author Note

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

The authors want to thank Dr. Frans Prins (Open University of the Netherlands, Educational Technology Expertise Center) for his useful contributions to this article

## Footnotes

1. A definition that corresponds to our view of competency-based assessment is given by Cizek (1997): (1) the planned process of gathering and synthesizing information relevant to the purposes of (a) discovering and documenting students' strengths and weaknesses, (b) planning and enhancing instruction, or (c) evaluating progress and making decisions about students, (2) the process, instrument or method used to gather the information. (p. 10).

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

Table/Figure Caption

Table 1

*Short Description of the Ten Quality Criteria for CAPs*

| Criterion | Short description |
|---|---|
| Authenticity | The degree of resemblance of a CAP to the criterion situation, usually those competencies needed in the future workplace. Gulikers et al. (2004) distinguish five dimensions that can vary in authenticity: the assessment task, the physical context, the social context, the assessment result or form, and the assessment criteria. |
| Cognitive complexity | The thinking processes and the fact that the assessment tasks should reflect the presence and level of required higher cognitive skills (Hambleton, 1996; Linn et al., 1991). |
| Comparability | CAPs should be conducted in a consistent and responsible way. The conditions under which the assessment is carried out should be, as much as possible, the same for all learners, scoring should occur in a consistent way, and large sampling across the content and situations of the competency at stake is necessary. |
| Costs and efficiency | The time and resources needed to develop and carry out the CAP, compared to the benefits. Evidence needs to be found that the (additional) investments in time and resources are justified by the positive effects, such as improvements in learning and teaching (Hambleton, 1996). |
| Directness | The degree to which teachers or assessors can immediately judge whether a student can function in a certain profession, without having to deduce or infer this. For example, using a student's reflections on how (s)he handles a specific situation does not directly show how (s)he deals with stress or unexpected situations. This can only be inferred. |

| | |
|---|---|
| Educational consequences | The intended, unintended, positive and negative effects of a CAP on learning and instruction (Linn et al., 1991; Messick, 1994; Schuwirth & Van der Vleuten, 2004). |
| Fairness | CAPs should not show bias to certain groups of learners and reflect the knowledge, skills and attitudes at stake, excluding irrelevant variance (Hambleton, 1996; Linn et al., 1991). |
| Meaningfulness | CAPs should have a significant value for both teachers and learners (Hambleton, 1996; Messick, 1994). To this, the value of the CAP in the eyes of the future employers and society as a whole could be added. |
| Reproducibility of decisions | The decisions made on the basis of the results of CAP should be accurate and constant over situations and assessors. Decisions should not depend on the assessor or the specific assessment situation. |
| Transparency | CAPs should be clear and understandable to all stakeholders. Learners should know the scoring criteria, who the assessors are and what the purpose of the assessment is. External controlling agencies should be able to get a clear picture of the way in which a CAP is developed and carried out. |

Table 2

*Scales of the Questionnaire filled out by the teachers*

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

Table 3

*Means and SD of criterion scales*

| | Overall | | Vocational education | | Pre-vocational education | | |
|---|---|---|---|---|---|---|---|
| Criterion scale | *M* | *SD* | *M* | *SD* | *M* | *SD* | Difference |
| Transparency | 4.50 | .59 | 4.54 | .57 | 4.35 | .65 | .19 |
| Authenticity | 4.13 | .66 | 4.17 | .65 | 4.00 | .72 | .17 |
| Cognitive complexity | 4.08 | .70 | 4.41 | .61 | 3.93 | .91 | .48** |
| *Thinking level* | | | | | | | |
| Cognitive complexity | 4.08 | .72 | 4.07 | .70 | 4.10 | .79 | -.03 |
| *Thinking process* | | | | | | | |
| Comparability | 4.08 | .87 | 4.10 | .86 | 3.99 | .92 | .11 |
| Meaningfulness | 4.05 | .67 | 4.04 | .66 | 4.10 | .69 | -.06 |
| Fairness | 4.04 | .66 | 4.08 | .65 | 3.85 | .67 | .23* |
| Costs & Efficiency | 4.04 | .80 | 4.13 | .77 | 3.68 | .84 | .45* |
| Educational | 4.03 | .71 | 4.05 | .71 | 3.95 | .71 | .10 |
| Consequences | | | | | | | |
| Directness | 3.93 | .74 | 3.96 | .74 | 3.79 | .70 | .17 |
| Reproducibility | 3.88 | .73 | 3.89 | .72 | 3.81 | .77 | .08 |

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

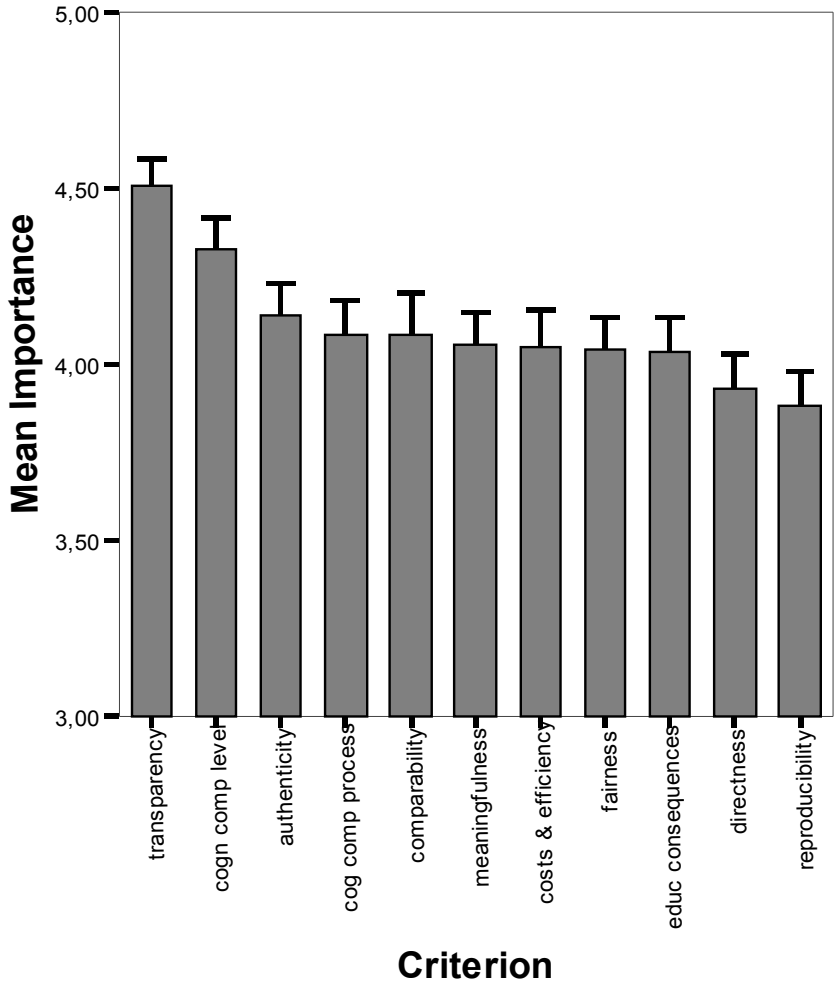
5,00
4,50
4,00
3,50
3,00
Mean Importance
transparency
cogn comp level
authenticity
cog comp process
comparability
meaningfulness
costs & efficiency
fairness
educ consequences
directness
reproducibility
Criterion

Figure 1.

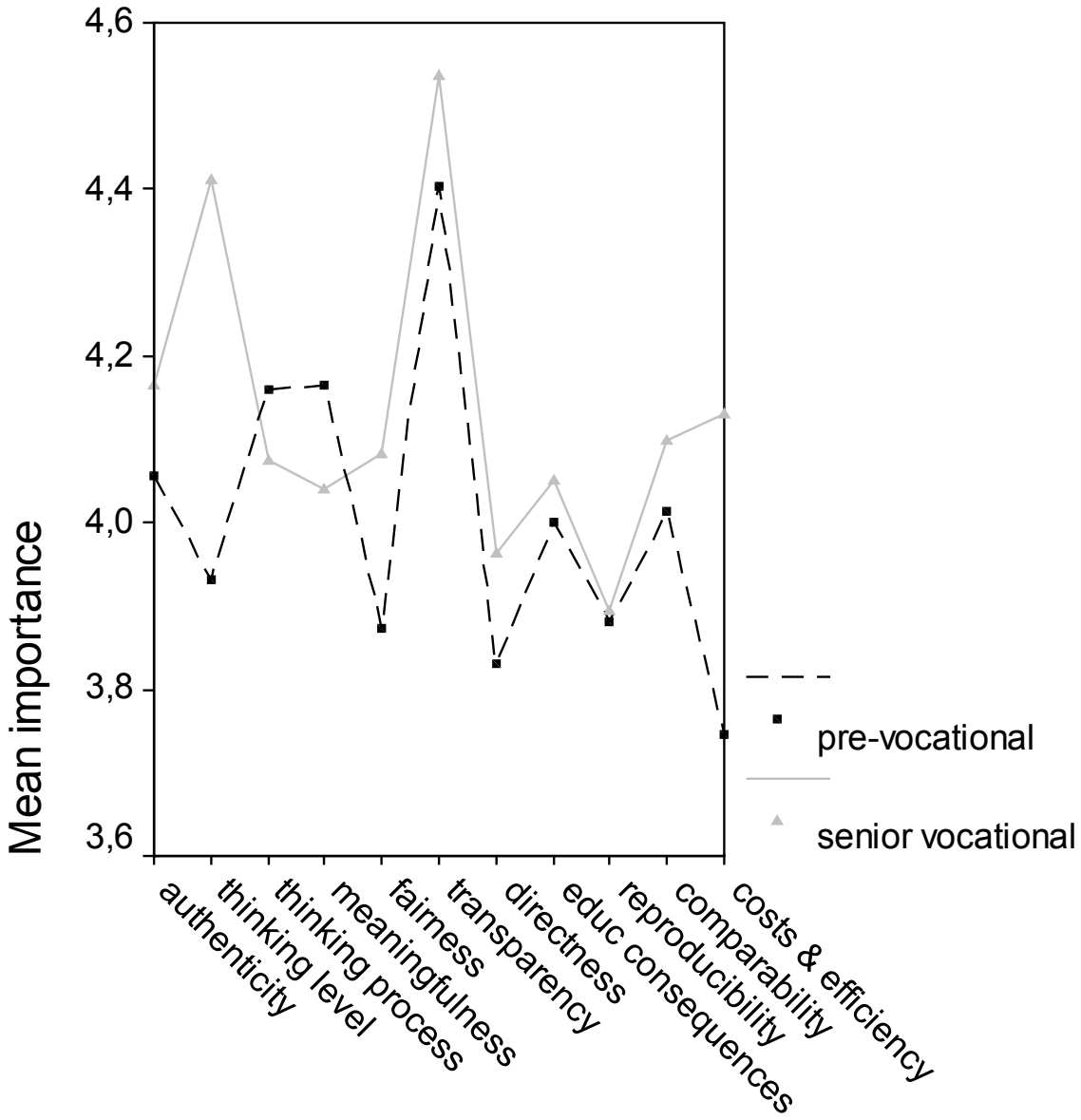
4,6
4,4
4,2
4,0
3,8
3,6
Mean importance
pre-vocational
senior vocational
authenticity
thinking level
thinking process
meaningfulness
fairness
transparency
directness
educ consequences
reproducibility
comparability
costs & efficiency

Figure 2.