Determining the quality of competences assessment programs:

1Running head: CAP QUALITY SELF-EVALUATION

This article was published as

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

Determining the Quality of Competences Assessment Programs: A Self-Evaluation Procedure

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

[a]Utrecht University, the Netherlands

[b]Open University of the Netherlands

[c]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, Utrecht University, Department of Educational Sciences, P.O. Box 80140, 3508 TC, Utrecht, The Netherlands. E-mail: L.K.J.Baartman@uu.nl

Abstract

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

Keywords: program evaluation, alternative assessment, evaluation criteria, self evaluation

Determining the Quality of Competences Assessment Programs: A Self-Evaluation Procedure

Education is undergoing fundamental changes in many European countries. In the Netherlands, new qualification structures for vocational education have been developed which are based on competences and work-related experiences. The rationale behind these innovations is to better link educational programs to job requirements and to enable vocational education to incorporate new developments in the market-place (Tillema, Kessels, & Meijers, 2000). From 2008 on, Dutch vocational institutions are legally bound to adopt a competence-based curriculum, focusing on the competences (knowledge, skills and attitudes) needed in relevant job situations. As an important part of education, assessment is changing as well (Birenbaum, 1996; Dochy & McDowell, 1997). Competence-based curricula require different assessment approaches to adequately determine competence-acquisition. As competence can be seen as the capacity to enact specific combinations of knowledge, skills and attitudes in appropriate job contexts (Lizzio & Wilson, 2004), assessment should focus on the integration of these three elements. This implies that in addition to assessing content knowledge, skills and attitudes should be assessed, and this should be done in an integrated way.

In the transition towards assessment of competence, assessment quality has played a key role. Traditional knowledge-focused assessment approaches are currently being criticized by a number of researchers. Recently, Birenbaum et al. (2006) stated that traditional assessment approaches focus on assessment *of* learning instead of assessment *for* learning, are limited in scope, and ignore individual differences increasingly encountered in education. Although part of this might be true, alternative assessment approaches currently being developed are not without problems either. Though they are supposed to be more valid than classical assessments (Linn, Baker, & Dunbar, 1991; Birenbaum, 1996), some feel that the evidence against classical tests is not as strong as has been claimed, and that the claim that newer forms of assessment are more

valid and suitable still needs empirical confirmation (e.g., Glaser & Silver, 1994; Hambleton & Murphy, 1992; Messick, 1994). This article does not attempt to resolve the dispute between traditional and new approaches to assessment. Instead, we argue that (1) it is unwise to assume that new approaches to assessment are a panacea for solving all assessment problems, and (2) that traditional and new assessment can be viewed as playing complementary rather than contradictory roles (Birenbaum, 1996; Maclellan, 2004). Therefore, in earlier publications (Baartman, Bastiaens, Kirschner, & Van der Vleuten, 2006) we proposed the use of Competence Assessment Programs (CAPs), which are defined as combinations of traditional and new forms of assessment in an assessment program, which can have both formative and summative functions (Baartman, Bastiaens, Kirschner, & Van der Vleuten, 2006).

Program Quality versus Single Method Quality

There are a number of reasons why it is important to think in terms of programs of assessment and why the quality of such a program should be evaluated as a whole. First, since competences involve the integrated application of knowledge, skills and attitudes, it is often argued that one single assessment method is not enough to assess competences and that a mix of methods should be used instead (e.g. Chester, 2003; Van der Vleuten & Schuwirth, 2005). Second, Knight (2000) argues for a program-wide approach to assessment in which attention is concentrated upon all assessment arrangements in complete educational programs. The advantage of this approach is that the reliability pressure on low stakes assessments in a program can be reduced and the resources freed up can be invested in the development of costly and reliable (and more valid!) assessments where they are needed, namely in high stakes situations. Third, a CAP comprises assessments with both formative and summative purposes. The main functions of formative assessment are providing feedback and generating appropriate learning activities, whereas summative assessment mainly serves to enable grading decisions (Black &

Wiliam, 1998; Gibbs, 1999). Knight (2000) argues for the need to make an explicit distinction between formative and summative assessment, in which reliability is less important for formative assessment and where summative assessments should be made as reliable as possible. Although he does urge not to diminish the validity of summative assessments, we feel that making such a clear distinction between formative and summative assessments runs the risk of evaluating formative assessment on new, learning-related criteria, and summative assessment on traditional, technical criteria. As summative assessments also have a "formative potential" (Hickey, Zuiker, Taasoobshirazi, Schafer, & Michael, 2006) in steering students' learning processes, we argue that learning-related quality criteria are just as important for summative assessments and that CAP quality should be evaluated integrally.

To evaluate CAPs, this article uses 12 quality criteria developed and validated in earlier studies (Baartman et al., 2006; in press). Table 1 lists the quality criteria and gives a short summary of each. The rationale behind the quality criteria is that since CAPs consist of both classical and new forms of assessment, both traditional and new quality criteria are needed to evaluate their quality. Our previous work (Baartman et al., 2006; in press) addresses some problems with regard to the use of reliability and validity for CAPs and suggests operationalizing reliability and validity in a different way and complementing them with other quality criteria proposed for new forms of assessment, such as the consequences, meaningfulness and cognitive complexity of an assessment (e.g., Kane, 1992, 2004; Linn et al., 1991; Van der Vleuten & Schuwirth, 2005).

- Insert Table 1 about here –

Very little is known about how to determine the quality of an assessment program instead of the quality of a single assessment method. Stokking et al. (2004) state that the criteria used should depend on whether the assessment is used formatively or summatively. For formative

assessments, comparability and reproducibility can get less priority, whereas efficiency is very important to assure that feedback can be given often and efficiently. For summative assessments, special measures to assure comparability, reproducibility and fairness should be a standard procedure. Transferring these ideas to program quality implies that not all single assessment methods in a CAP must meet all quality criteria. Although we can be more lenient with regard to reproducibility and comparability for formative assessment, the summative assessments within a program should comply with all quality criteria, including learning- and feedback-related ones like meaningfulness and educational consequences. A CAP as a whole has to comply with all quality criteria. For example, high scores on authenticity cannot offset major deficits in cognitive complexity.

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

School Self-Evaluations

Assessment quality can be demonstrated in a large number of ways, of which self-evaluation is just one. Jonsson and Baartman (2006), for example, evaluated an assessment program by means of analyses of student examination scores and student questionnaires, and in many countries external auditing is a commonly used method to assure assessment quality. A self-evaluation method was chosen here because in many European countries, school self-evaluation is becoming an increasingly important approach to both school improvement and

accountability (McNamara & O'Hara, 2005). School self-evaluation or internal evaluation is carried out by a school itself, for example by a group of teachers, the department or school manager, a specific staff member, or a combination thereof. In contrast, external evaluation is carried out by someone outside the school, usually inspectors or governmental organizations, and mainly serves accountability purposes (Nevo, 1994, 2001). In these discussions about internal and external evaluation, school improvement and self-evaluation refer to the educational process as a whole, and not specifically to assessment.

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

Studies on the use of self-evaluation have shown positive results (e.g., McNamara & O'Hara, 2005; Nevo, 1994, 2001). Teachers appear to be willing to be self-critical and

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

In sum, research on self-evaluation shows the merits of self-evaluation, but also some possible pitfalls, one of which is the fact that schools often do not support their claims by using appropriate pieces of evidence. Previous studies, though, have not looked into the exact nature of the support presented in self-evaluations. This study does look at this support from the perspective of argumentation theory and takes a qualitative approach to gain a deeper understanding of the processes taking place during self-evaluation. Research on argumentation shows that the ability to provide support for one's claims cannot be taken for granted (Kuhn, 1991). If self-evaluation is to be a valuable approach to both school improvement and accountability, a precondition is that schools are capable of performing self-evaluations. In this study, a self-evaluation procedure was developed to assist schools in evaluating their newly developed CAPs, based on the 12 quality criteria for CAPs developed in earlier studies (Baartman et al., 2006, in press). Eight vocational schools participated. In each of these schools, three functionaries collaboratively evaluated their CAP using the self-evaluation procedure. It

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

The CAP Quality Self-Evaluation Procedure

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

Phase 1: Individual CAP Self-Evaluations

The individual self-evaluations of a school's CAP are carried out with a web-based evaluation tool, based on the twelve quality criteria. Before evaluating their CAP, the evaluators are asked to describe it by indicating the year(s) and level of education, and the assessment methods included. Examples of methods are given, including multiple choice test, written test with open questions, presentation, assessment of products made, assessment interview, criterion-

based interview, observation in a simulated situation, observation in the workplace, portfolio and proof of competence. Additional forms of assessment can be added by the user.

Subsequently, they evaluate their CAP on the 12 quality criteria for CAPs developed earlier (Baartman et al., 2006, in press). For a more elaborate description and discussion of the criteria we refer to our earlier studies. For the self-evaluation tool, these quality criteria are operationalized as indicators: more concrete aspects of a quality criterion in practice, though not too detailed that they turn the self-evaluation into just ticking off a checklist. Per quality criterion, four to six indicators are formulated, based on a literature study and an earlier carried out pilot study (e.g., Baartman et al., 2006, in press; Baume, Yorke, & Coffey, 2004; Benett, 1993; Dierick & Dochy, 2001; Dochy, Gijbels, & Van de Watering, 2004; Gulikers, Bastiaens, & Kirschner, 2004; Linn et al., 1991; McLellan, 2004; Miller & Linn, 2000; Moss, 1994; Schuwirth & Van der Vleuten, 2004). Along with the pre-determined indicators, two open fields are included for each criterion, so that users can include more and other indicators relevant to their situation. Table 2 gives an overview of all quality criteria and an abbreviated version of their indicators.

The CAP is evaluated both quantitatively and qualitatively. For the quantitative evaluation, the CAP self-evaluation tool asks the evaluators to rate the CAP on each indicator via an analog slide-bar that can be moved from “not at all” to “completely” (see Figure 1). An option “don’t know” was available as well. Behind this slide bar is a rating scale ranging from 0 to 100, which is invisible so as not to give evaluators the idea of giving a score or mark to their CAP. For the qualitative evaluation, the tool asks for support of the ratings given in the form of an example or evidence showing that the CAP indeed complies with the indicator. The self-evaluation tool is complemented by an instruction page and a vocabulary list in which all different assessment methods are defined and explained. The instructions and vocabulary list can

be accessed at any time. Figure 1 presents a screen dump of a page of the CAP self-evaluation tool.

- Insert Figure 1 about here -

Phase 2: Group Interview

After the individual CAP self-evaluations, all individual ratings and support thereof are assembled and collected in an overview of the school's CAP-quality. For each quality criterion, the overview presents the ratings and support of the indicators given by all evaluators. The overview is used as input for the group interview, which is meant to stimulate discussion and reflection on CAP quality and to result in an overall picture of the quality of the CAP evaluated. The group interview lasts about two hours and has a semi-structured character. First, the evaluators are asked to globally describe their CAP. They are given the list of assessment methods they ticked off in the self-evaluation tool and are asked to describe them more elaborately and to indicate the percentage of total assessment time devoted to each. Second, the overview with all evaluators' ratings and support is discussed and they are explicitly encouraged to comment on their own and each others' ratings and support. If they change their minds about a rating or support thereof during the group interview, they are allowed to adjust their initial rating and / or support given in the individual self-evaluation (comparable to a Delphi-study approach). This is noted down by the interviewer, who asks for further information or explanation if the:

a. Argumentation is unclear to the interviewer;
b. Interviewer thinks the argumentation is too weak to support the rating;
c. Evaluators have clearly different opinions. To get an indication of "a clearly different opinion" the range of ratings was divided into three categories: 0-35 (low), 36-65 (medium), and 66-100 (high). A clearly different opinion was operationalized as falling in different categories and differing at least 20 points.

To conclude the group interview, the evaluators are asked to collaboratively summarize the strong and weak aspects of their CAP, based on the individual self-evaluations and the group interview.

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

Method

Context of the Studies

The study was carried out in Laboratory Technology Education in upper secondary vocational institutions in the Netherlands. Within the Dutch educational system, after leaving primary schools, all pupils are required to enter secondary education where they can choose between general secondary education which leads to entrance to a university or polytechnic, and pre-vocational education (age 12-15). Pre-vocational education serves as a preparation for upper

secondary vocational education (age 15-18). Laboratory Technology is a vocational program preparing students for a job as laboratory assistant or laboratory technician. The schools participating in this study were organized in a national consortium of vocational schools that started to implement problem-based education in 2000/2001 and is now working towards competence-based education.

Participants

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

Procedure

All schools were contacted through the national consortium of vocational schools, within which the Laboratory Schools are organized as a content-specific working group. One week before the group interview, all participants received an email asking them to independently fill out the CAP quality meter. The three participants from the same school were asked to first collaboratively determine the CAP they would evaluate, for example all assessments used in the first year of the educational program. This ensured all participants from one school had the same CAP in mind. The participants then individually used the CAP self-evaluation tool to evaluate

the chosen CAP. The group interviews were carried out by the first author approximately a week later and lasted about 2 hours. At the start of the group interview, the participants were presented with the overview of all individual CAP self-evaluations. All interviews were audio taped with permission of the participants.

Data analyses

To answer the first research question on the contribution of the individual phase and the group interview to the processes and outcomes of the self-evaluation, both quantitative and qualitative analyses were carried out. First, the percentage of ratings completed with a piece of support was calculated before and after the interview. The ratings given were divided into the three categories used in the group interview: low (0-35), medium (36-65), and high (66-100), together with a "don't know" category. The percentages of low, medium, high ratings per indicator given before and after the interview, and the changes made during the interview (e.g. a change from a low rating to a high rating) were calculated. Second, all group interviews were transcribed literally and analyzed qualitatively. The first author analyzed the group interviews by noting recurrent themes, for example if the participants had thought of a specific part of their CAP instead of the entire CAP when giving their ratings and support. The first themes were identified when analyzing the first interviews. The other interviews were used to check whether they could be found again and new and other themes were added to the first ones. This process continued until no new themes could be identified, which was the case after analyzing all interviews three times. Then, the list of themes found by the first author was given to a researcher not involved in the current project, who independently analyzed the group interviews. She identified the themes listed by the first author by marking the parts of the transcribed interviews belonging to each theme and added new themes to the list made by the first author.

The first author and the independent researcher discussed the themes until agreement over the list was reached.

To answer the second research question on the nature and quality of the support, Miles and Huberman's (1984) phases for qualitative data analysis were followed, in which qualitative data are first meaningfully reduced or reconfigured (data reduction), then organized into different data displays such as diagrams and matrices (data display), from which conclusions can be drawn and verified in the last phase (conclusion and verification). In the first phase of analysis, a summarizing display was constructed for each school with the ratings and support given in the individual CAP self-evaluations, complemented with the ones found in the typed out group interviews. The support was summarized over the three participants per school, resulting in an overview of the support given for each school. If the participants agreed on the support given, this was summarized in the overview. If they did not agree, two or three different pieces of support were included in the analysis. For the second phase of analysis, the eight overviews for the separate schools were assembled in a meta-matrix. The support was now summarized over schools, resulting in a so-called ordered matrix including all different pieces of support together with their ratings, which were again categorized into low, medium and high. From this ordered meta-matrix, conclusions were drawn in the last phase of analysis. To assure the qualitative analyses did not depend on the authors' personal and subjective interpretations (verification), a check was carried out by a researcher not involved in the current project who independently re-constructed the data displays. Differences were discussed and changed in accordance with both researchers' opinions. To assure further verification, the first and second author together carried out the final conclusion phase, for which a flow chart for coding the quality of arguments developed by Clark and Sampson (2005) was adjusted for this research. Clark and Sampson's flow chart is based on Toulmin's (1958) well-known scheme of the layout of arguments. In

argumentation literature, some researchers analyze the quality of argumentation by investigating if every element of Toulmin's scheme is present (e.g. Simon, Erduran, & Osborne, 2006), but Clark and Sampson argue that these analyses should also include judgments of the quality of the arguments, and not just their absence or presence. The flow chart classifies the quality of argument as either: no support (level 0), using explanation as support (level 1), using evidence as support (level 2) and coordinating multiple pieces of evidence or multiple connections between ideas in the evidence (level 3). The flow chart was adjusted by referring to the quality of CAPs instead of arguments in a group discussion. Figure 2 presents the adjusted flow chart used in this research. The first and the second author independently coded all support using the flow chart and kappa values were calculated to check for interrater reliability. After coding the argumentations of two quality criteria, the initial interrater reliability was found to be mediocre to good (.51 and .70). The different codings were discussed and the largest differences between the two researchers appeared to involve the distinction between level 0 and level 1. From some pieces of support, it did not become completely clear whether the participant was really adding any new information, or was merely repeating the indicator. It was decided upon to score the indicator as level 0 when it was not completely clear what the participant was exactly referring to and whether this could be considered as additional information to the indicator, although additional information might have been present implicitly. After resolving these differences, all support was scored and interrater reliabilities were found to be satisfactory (Cohen's kappa ranging from .70 to .87). The codes from the second author were used for further analyses, as the first author conducted all interviews and could thus be more biased towards certain schools.

Results

Before presenting the results with regard to the two research questions posed, the Cronbach's Alpha values of the criterion scales of the 12 quality criteria are discussed here.

Although we did include the possibility to include other indicators than the ones proposed and we do not pretend to give a full overview of all possible indicators, the indicators were designed as a scale of each quality criterion. Table 2 presents the Cronbach's Alpha values found for the criterion scales (in bold) and the item-total correlation for each indicator. Taking .60 as an acceptable alpha value, six criteria initially could not be considered as a scale. In addition, a number of indicators had low item-total correlations. These correlations should be higher than . 35, but lower values are accepted if items cannot be missed theoretically.

A reason that may explain some low Alpha values is that Cronbach's Alpha increases when the sample size increases and when the number of items within a scale increases. In this study, we had a relatively small number of participants (N = 22), and some indicators had many missing values. In this case, the missing values are the percentages in the "don't know" category in table 2. As any unclear indicators were explained during the group interview, a high percentage "don't know" seems to indicate that the participants indeed did not know whether their CAP complied with the indicator or not, and not whether they understood the indicator or not. For the Alpha values this resulted in a lower sample size, which was sometimes reduced by almost half. For this reason, the Alpha values were re-calculated using mean substitution for the criteria with an insufficient Alpha value. This resulted in an acceptable Alpha value for the quality criterion acceptability. The other quality criteria still had insufficient Alpha values, which necessitated the deletion of indicators. The last column of Table 2 shows the re-calculated Alpha values with deletion of the indicators with the lowest item-total correlations. The re-calculated Alphas are sufficient, although the reproducibility and transparency scales need to be interpreted with some caution. Although statistically the deletion of indicators was necessary, at this first stage of development and use of the indicators and scales we are reluctant towards permanently deleting indicators. At this stage of development of CAPs it is very well possible that schools do

pay attention to one indicator, but not to another. For example, for acceptability some schools may have asked students' opinions, but not employers', and other schools may have done so the other way around. This difference results in a low Alpha value, which does not mean that the theoretical concepts within the scale do not fit together. In sum, at this moment we had to delete some indicators from the scales and some indicators indeed may not fit in a scale theoretically, but further research using larger samples is needed before final conclusions can be reached here. Moreover, some indicators may get less "don't know" answers in the future, when schools are more used to the newer ideas of the quality of assessment presented in the indicators. At this moment, we will present the results of this study on the scale level as much as possible, but we will refer to the indicators when necessary.

The Individual Self-Evaluations and Group Interviews

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

Ratings and support before and after the group interview

The first two columns of Table 2 present the percentages of ratings completed with a piece of support before and after the interview. As can be seen, before the group interview 63% of the ratings were supported with a piece of evidence. After the interview this percentage had increased to 76%, meaning that support was added or complemented during the interview. As the data are non-parametric, this difference was tested by means of a Wilcoxon's signed ranks test and was found to be significant ($Z = -6.182$, $p < .001$). In total, 58 ratings were changed during the group interview. Most changes were from a low rating to a high rating (15), followed by

changes from a low to a medium rating (9) and a low rating to a "don't know" (9). Apparently, the group interview and the discussions between the participants caused them to give higher ratings than they had initially given individually before the interview. The changes from a low rating to a "don't know" were mostly caused by the fact that the participants realized they had given a rating without being able to support it: "Actually, I don't have any experience with assessing choices at this moment … I should have put 'don't know'" [school 1]. It needs to be remarked here that in total very few changes in ratings were made during the interview. In total 1254 ratings were given, of which 58 were changed (4.6%). Apparently, the interviews had a greater effect on the support than on the ratings given.

- **Insert Table 2 about here –**

In addition, Table 2 presents the percentages of low, medium and high ratings given after the interview. The ratings after the interview were taken here to include any "corrections" made and because few changes were made at all during the interview. In total, many more high (M = 50%) than low (M = 18%), medium (M = 20%) and N-ratings (M = 12%) were given. Friedman's non-parametric test showed that the differences between the percentages of low, medium and high ratings was significant ($\chi^2 = 72.727$, $p < .001$). Wilcoxon's signed ranks tests showed that the differences between the number of high and low ratings and the difference between the number of high and medium ratings were significant (Z = -5.501, p = .000 and Z = -6.142, $p < .001$ respectively). The difference between the number of low and medium ratings was found to be non-significant (Z = -1.453, p = .146). Apparently, the participants gave their CAP relatively high ratings. The highest percentage of high ratings was found for Comparability (85%). This is a quality criterion that traditionally has been paid much attention to, and this does not seem to have decreased during the transition towards competence-based education. The lowest percentages of high ratings were found for Meaningfulness (33%), Costs & Efficiency

(34%), and Cognitive Complexity (36%). These quality criteria are newer and schools may be less familiar with these concepts.

Recurrent themes in the group interview

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

Rating and supporting the indicators:

(1) The participants give ratings and support for a broader CAP than agreed upon;

(2) The participants give ratings and support for a specific smaller part of the CAP;

(3) The participants describe how they would like their CAP to be, instead of rating and supporting the actual situation;

(4) The participants say their school is in a transition period towards competency-based education, and therefore some indicators cannot be answered yet or will change in the near future;

The added value of the group interview:

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

(6) Caused by the self-evaluation process and the discussion in the group interview, the participants come up with spontaneous ideas for improving their CAP;

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

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

Recurrent themes: Rating and supporting the indicators

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

experience with competence-based education and corresponding CAPs.

Recurrent themes: Added value of the group interview

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

Recurrent themes: Formative versus summative assessment and the EQC

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

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

Nature and Quality of Support

Figure 3 presents the percentage of support coded at each level of argumentation distinguished by Clark and Sampson (2005). As can be seen, the main part (in total 56%) of the support was coded as level 1 (explanation as support). Argumentation level 0 (no argumentation) was assigned to 22% of the support and 23% of the support was coded as level 2 (evidence as support). Level 3 (coordinating multiple pieces of evidence) was not found in our data. Support at level 0 was mainly characterized by the fact that they were irrelevant to the indicator at stake or that they were merely a repetition of the indicator. One example is a participant responding to the Comparability-indicator "Assessment procedure comparable" by saying that "We try to assure all assessment procedures are comparable" [school 3]. Some support at level 0 was characterized by the fact that the participant gave his or her opinion on the matter instead of providing evidence. For example, reacting to the indicator "Assessors with different background", one participant reacted "I think the teacher should do the assessments. Students should not have any influence on this" [school 7]. Argumentation level 1 was mainly characterized by participants presenting their own personal experiences, like this participant does for the indicator "Giving and receiving feedback": "I experience the self-assessment generates valuable feedback on the student's strengths and weaknesses" [school 3]. Support at level 2 involved actual pieces of evidence, for example for the indicator "Improved if negative effects" one manager remarked: "We recently conducted an evaluation of the assessment and did a brainstorm session with the teachers. We formulated the weaknesses and little groups of teachers are now trying to find solutions to this, for example about how to give better and more immediate feedback" [school 7].

Conclusions and Discussion

The purpose of this study was to explore whether schools are capable of evaluating their own competences assessment program or CAP. A CAP self-evaluation procedure was developed

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

CAP being evaluated.

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

This study had an exploratory character and focused on the process of carrying out the CAP self-evaluation, and not on the final product of this self-evaluation, that is the actual quality of the CAP being evaluated. Although this a very interesting and important question that will be addressed in further analyses and studies, we think it is important to first focus on the process of the self-evaluation. Both the idea of carrying out self-evaluations instead of external evaluations

and the idea of evaluating programs of assessment instead of single assessment methods are relatively new. Future research is still needed here. For example, for formative purposes, further research is needed to explore whether the CAP self-evaluation procedure indeed, as it seems to do, stimulates reflection on CAP quality and if this leads to future improvements of the CAP. With regard to program quality as opposed to single assessment method quality, there still is a need for more explicit standards specifying acceptable minimum levels of all quality criteria for the program as a whole. These standards are necessary for summative evaluation of assessment programs, but can also serve as a point of reference or benchmark for schools when carrying out (formative or summative) self-evaluations.

For now, we conclude that the evaluation of assessment programs by means of a self-evaluation procedure seems to be possible for formative purposes, but schools need to be supported in the process. A group interview guided by an expert in the field of assessment quality seems necessary to get an overall picture of the CAP's quality. For summative purposes and accountability, though, issues of the reliability of self-ratings become more important, and more research is needed on this matter. The combination of formative and summative purposes of self-evaluation, as is done when self-evaluation is used for both school improvement and accountability, could cause problems in this respect. Differences between judges are generally unwanted in summative evaluation, whereas they may be beneficial for formative purposes by helping generate discussion and stimulate reflection. Finally, with regard to the evaluation of CAP quality, the integral framework of the twelve quality criteria used here seems to be promising to evaluate program quality in an integrated way.

References

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

Baartman, L. K. J., Bastiaens, T. J., Kirschner, P. A., & Van der Vleuten, C. P. M. (2006). The wheel of competency assessment. Presenting quality criteria for Competency Assessment Programs. Studies in Educational Evaluation, 32, 153-177.

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

Baume, D., Yorke, M., & Coffey, M. (2004). What is happening when we assess, and how can we use our understanding of this to improve assessment? Assessment and Evaluation in Higher Education, 29, 451-477.

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

Birenbaum, M. (1996). Assessment 2000: Towards a pluralistic approach to assessment. In M. Birenbaum & F. J. R. C. Dochy (Eds.), Alternatives in assessment of achievement, learning processes and prior knowledge (pp. 3-29). Boston, MA: Kluwer Academic Publishers.

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

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

Clark, D. B. & Sampson, V. D. (2005). Analyzing the quality of argumentation supported by personally-seeded discussions. In T. Koschman, T. Chan & D. D. Suthers (Eds.), Computer-

supported collaborative learning 2005: the next 10 years! (pp. 76-85). Tapei, Taiwan: Lawrence Erlbaum Associates.

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

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

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

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

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

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

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

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

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

Kingdom.

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

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

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

Kuhn, D. (1994). The Skills of Argument. Cambridge, England: Cambridge University Press.

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

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

Lizzio, A. & Wilson, K. (2004). Action learning in higher education: an investigation of its potential to develop professional capability. Studies in Higher Education, 29, 469-488.

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

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

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

Miles, M.B., & Huberman, A.M. (1984). Qualitative Data Analysis. A Sourcebook of New Methods. Beverly Hills, CA: Sage Publications.

Miller, M.D., & Linn, R.L. (2000). Validation of performance-based assessments.

Applied Psychological Measurement, 24, 367-378.

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

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

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

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

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

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

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

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

Tillema, H. H., Kessels, J. W. M., & Meijers, F. (2000). Competencies as building blocks for integrating assessment with instruction in vocational education: A case from The Netherlands. Assessment & Evaluation in Higher Education, 25, 265-278.

Toulmin, S. (1958). The Uses of Argument. Cambridge, England: Cambridge University Press.

Valli, L, & Rennert-Ariev, P. (2002). New standards and assessments ? Curriculum

transformation in teacher education. Journal of Curriculum Studies, 34, 201-225.

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

Table 1

Short Description of the Twelve Quality Criteria for CAPs

Criterion	Short description
Acceptability	All stakeholders (e.g. students, teachers, employers) should approve of the assessment criteria and the way the CAP is carried out. They should have confidence in the CAP's quality
Authenticity	The degree of resemblance of a CAP to the future workplace. Gulikers, Bastiaens and Kirschner (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	A CAP should reflect the presence of the cognitive skills needed and should enable the judgment of thinking processes
Comparability	CAPs should be conducted in a consistent and responsible way. The tasks, criteria and working conditions should be consistent with respect to key features of interest
Costs and efficiency	The time and resources needed to develop and carry out the CAP, compared to the benefits
Educational consequences	The degree to which the CAP yields positive effects on learning and instruction, and the degree to which negative effects are minimized
Fairness	Students should get a fair chance to demonstrate their competences, for example by letting them express themselves in different ways and making sure the assessors do not show biases
Fitness for Purpose	Alignment among standards, curriculum, instruction and assessment. The assessment goals and methods used should be compatible with the educational goals
Fitness for Self-Assessment	CAPs should stimulate self-regulated learning of students. CAPs should include specific methods to foster such learning such as practice in self-assessment and giving and receiving feedback
Meaningfulness	CAPs should have a significant value for all stakeholders involved (e.g.

	students, teachers, employers).
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 (Van der Vleuten & Schuwirth, 2005).
Transparency	CAPs should be clear and understandable to all stakeholders (e.g. students, teachers, employers). 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

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

Criteria and Indicators	Before interview % Subst.	After interview % Subst.	% low ratings (0-35)	% med ratings (36-65)	% high ratings (65-100)	% don't know	α & Item-Total	re-cal. α & Item-Total
Acceptability	**66**	**78**	**11**	**21**	**43**	**25**	**.51**	**.65**
1 Students approve of criteria	59	82	9	14	59	18	.28	.44
2 Students approve of procedure	55	73	9	32	41	18	.87	.43
3 Teachers approve of CAP	73	82	14	27	55	5	-.56	.27
4 Employers approve of CAP	55	86	9	14	18	59	.38	.25
5 Confidence in quality CAP	59	68	14	18	41	27	.61	.69
Authenticity	**76**	**85**	**15**	**23**	**61**	**1**	**.70**	
1 Assessment tasks resemble job	82	82	0	9	91	0	.23	
2 Working conditions resemble job	82	91	18	41	41	0	.56	
3 Social context resembles job	68	82	27	23	50	0	.64	
4 Assessment criteria resemble job	73	86	14	18	64	5	.55	
Cognitive complexity	**65**	**70**	**30**	**24**	**36**	**10**	**.74**	
1 Tasks trigger thinking steps	68	73	14	27	41	18	.64	
2 Explain choices	68	73	41	18	32	9	.47	
3 Criteria address thinking steps	55	59	41	23	23	14	.72	
4 Tasks require thinking level	68	77	23	27	50	0	.36	
Comparability	**65**	**72**	**5**	**8**	**85**	**2**	**.72**	
1 Assessment tasks comparable	77	86	9	9	82	0	.18	
2 Working conditions comparable	59	68	0	18	77	5	.65	
3 Assessment criteria comparable	64	64	0	5	91	5	.71	
4 Assessment procedure comparable	59	68	9	0	91	0	.59	
Costs & Efficiency	**56**	**69**	**26**	**18**	**34**	**22**	**.41**	**.69**
1 Time and money estimated	55	73	45	14	23	18	.44	.67
2 Deliberately choosing mix	55	68	32	14	36	18	.42	.69
3 Yearly evaluation of efficiency	59	73	18	23	45	14	.35	.23
4 Positive effects outweigh investments	55	64	9	23	32	36	-.25	
Educational Consequences	**65**	**64**	**18**	**22**	**42**	**18**	**.46**	**.71**
1 Desired learning processes stimulated	64	73	32	23	41	5	.63	.52
2 Positive influence on students	59	73	18	23	27	32	.25	.43
3 Positive influence on teachers	55	68	18	27	27	27	-.05	.57
4 Improved if negative effects	77	86	5	14	82	0	.36	
5 Curriculum adapted if CAP warrants	73	64	18	23	32	27	.17	.49
Fairness	**60**	**75**	**7**	**15**	**63**	**15**	**-.44**	**.77**
1 Procedures to rectify mistakes	59	73	0	18	59	23	.05	.57
2 Weights based on importance	68	82	32	14	41	14	-.42	
3 Assessors not prejudiced	59	77	5	23	64	9	.26	.81
4 Various types of assessment tasks	45	64	0	14	77	9	-.38	
5 Student think CAP is fair	68	77	0	5	73	23	.25	.51

Fitness for Purpose		**68**	**85**	**17**	**16**	**65**	**1**	**.70**	**.79**
1	Coverage of competence profile	77	95	0	23	77	0	.78	.78
2	Integrated assessment of K/S/A	77	95	41	27	32	0	.58	.64
3	Mix of different assessment forms	59	77	0	9	91	0	-.32	
4	Both summative and formative forms	64	77	32	0	68	0	.48	.50
5	Forms match with educational goals	64	82	14	23	59	5	.74	.72
Fitness for Self-Assessment		**61**	**69**	**31**	**23**	**40**	**7**	**.86**	
1	Self- and peer-assessment	73	95	18	27	55	0	.49	
2	Giving and receiving feedback	59	68	23	32	41	5	.68	
3	Reflection on personal development	55	55	32	18	41	9	.92	
4	Formulation of personal learning goals	59	59	50	14	23	14	.75	
Meaningfulness		**51**	**68**	**24**	**19**	**33**	**25**	**.93**	
1	Feedback formative useful	55	86	18	14	41	27	.83	
2	Feedback summative useful	64	95	23	23	27	27	.87	
3	Assessment is opportunity to learn	41	55	41	23	18	18	.81	
4	Students think criteria meaningful	41	50	23	18	27	32	.68	
5	Teachers/employers think criteria meaningful	55	55	14	18	50	18	.88	
Reproducibility of decisions		**63**	**85**	**27**	**27**	**40**	**6**	**.38**	**.59**
1	Several times	68	91	18	32	41	9	.04	
2	Several assessors	68	86	9	23	59	9	.36	.41
3	Assessors with different background	64	86	55	14	27	5	.13	.24
4	Equal discussion between assessors	64	86	23	18	50	9	.23	.41
5	Trained and competent assessors	55	86	32	45	23	0	.07	.28
6	Several work situation	59	73	23	32	41	5	.42	.46
Transparency		**65**	**76**	**9**	**25**	**57**	**9**	**.43**	**.58**
1	Student know formative of summative	77	86	0	23	73	5	.38	.47
2	Students know criteria	59	73	9	41	36	14	-.06	.36
3	Students know procedures	59	64	5	32	55	9	.49	.51
4	Teachers know and understand	73	86	0	27	68	5	.27	.26
5	Employers know and understand	64	82	23	9	50	18	.36	.21
6	External party can audit	59	64	18	18	59	5	.01	
Total		**63**	**76**	**18**	**20**	**50**	**12**		

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

Figure Caption

Figure 1. Screen dump of the CAP self-evaluation tool

Figure 2. Scheme for analyzing the nature and quality of support

Figure 3. Percentages of support within in the 3 levels of argumentation

Figure 1. Screen dump of the CAP self-evaluation tool

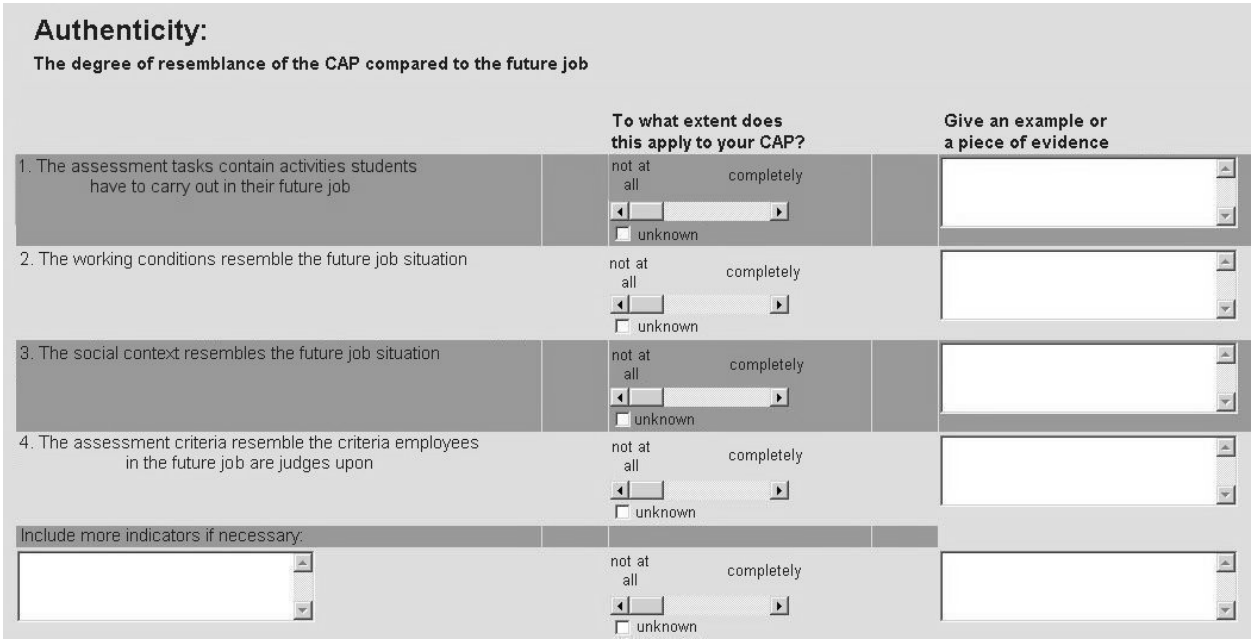

Figure 2. Scheme for analyzing the nature and quality of support

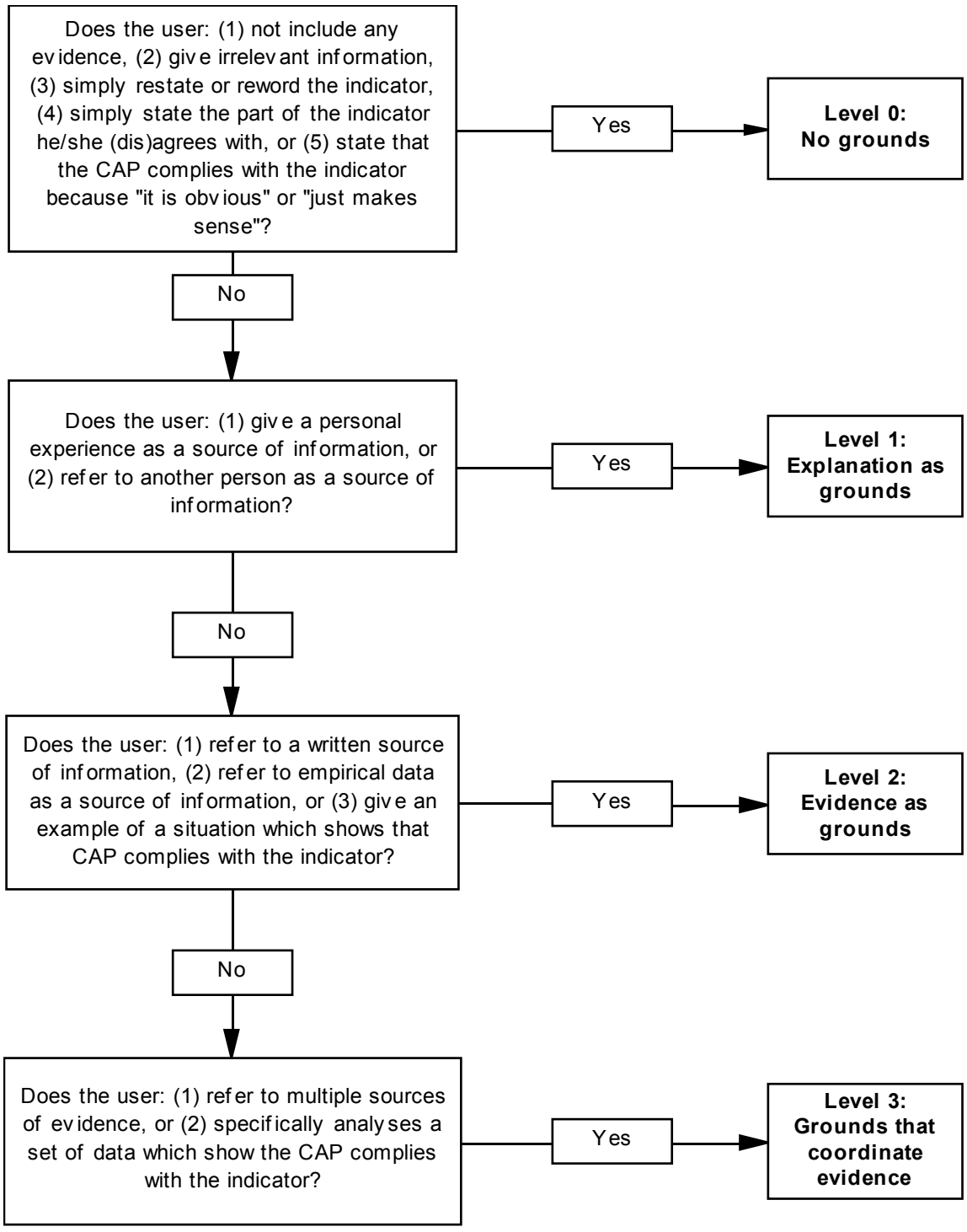

Figure 3. Percentages of support within in the 3 levels of argumentation