Assessment criteria for competency-based education: a study in nursing education

Instr Sci (2014) 42:971–994
DOI 10.1007/s11251-014-9326-5

Assessment criteria for competency-based education: a study in nursing education

Greet M. J. Fastré · Marcel R. van der Klink · Pauline Amsing-Smit · Jeroen J. G. van Merriënboer

Received: 15 July 2012 / Accepted: 26 June 2014 / Published online: 13 July 2014

Abstract This study examined the effects of type of assessment criteria (performance-based vs. competency-based), the relevance of assessment criteria (relevant criteria vs. all criteria), and their interaction on secondary vocational education students' performance and assessment skills. Students on three programmes in the domain of nursing and care ($N = 93$) participated in the study. Results show that students who were given the relevant criteria made more accurate assessments of an expert model, performed better on a test and achieved higher instructional efficiency (defined as the relationship between performance and mental effort) compared to students who were given all criteria. Students who were given performance-based assessment criteria made more accurate assessments of an expert model and scored higher on task performance during practice compared to students who were given competency-based assessment criteria. They invested less mental effort in the assessments, resulting in higher instructional efficiency. An interaction effect for the concreteness of answers shows that the combination of performance-based and relevant criteria leads to superior analysis of worked examples compared to the other combinations of criteria.

Keywords Assessment criteria · Competency-based · Performance-based · Relevance of criteria

G. M. J. Fastré · M. R. van der Klink · J. J. G. van Merriënboer
Centre for Learning Sciences and Technologies, Open University of the Netherlands, Heerlen, The Netherlands

G. M. J. Fastré · J. J. G. van Merriënboer
School of Health Professions Education, Maastricht University, Maastricht, The Netherlands

G. M. J. Fastré (✉)
Zonhovenstraat 85, 3500 Hasselt, Belgium
e-mail: greetfastre@hotmail.com

P. Amsing-Smit
Regional Education Centre A12, Ede, The Netherlands

Introduction

Both nursing students and car mechanic students must be able to analyse problems, cooperate with others and present information. The contexts in which they have to demonstrate these competencies, however, differ significantly. A student nurse must be able to analyse a patient file, whereas a car mechanic student must be able to analyse a car defect. Nevertheless, with the recent introduction of competency-based education in secondary vocational education (level 3 and 4 of the European Qualifications Framework) in the Netherlands, profiles have been composed for core tasks, work processes and competencies that are identical for all programmes, even though these programmes are aimed at training competent professionals who are well equipped to deal with future developments in very different fields (Biemans et al. 2004).

In recent decades, competency-based education has gained attention in vocational and higher education. This concept was embraced to improve the rigour and relevance of the curriculum, and to ensure that students not only memorize facts but are able to apply the knowledge to, for example, nursing practice (see Malone and Supri 2012).

Although on an aggregated level competency-based education always entails the relatedness of skills, knowledge and attitudes, it is applied differently in various parts of Europe. In England, competency-based vocational education is usually rooted in a functionalist view that refers to the performance of fragmented and narrowly defined tasks with a minimal underpinning of knowledge. This overall picture, however, is not observed in all education sectors. In nursing education, competency-based frameworks and assessments refer to the substantial knowledge needed for task performance. In France, on the other hand, competency is more closely linked to notions that emphasize the multidimensional nature, a stronger relation between practical and theoretical knowledge, and the contribution of personal and social qualities to task performance (see Brockmann et al. 2008). In the Netherlands, this latter, broader view on competency-based education can be observed (see van der Klink et al. 2007).

Although competencies differ because of the work context, a fixed set of 25 general competencies for level 4 competencies (e.g. 'working together' and 'consulting') has been developed. Students must have mastered these competencies by the end of all educational programmes in secondary vocational education (COLO 2008; Norcini et al. 2011). For each task, a subset of these competencies applies.

Assessment strongly influences what and how students learn. This implies that an innovation from knowledge-based to competency-based education, aimed at a better integration of knowledge, skills and attitudes (van Merriënboer and Kirschner 2007), can only be successful if the assessment programme is also competency-based (Birenbaum 2003). The main goal of competency-based assessment is to assess students' ability to perform professional tasks in accordance with specific criteria (Gulikers et al. 2010).

For novice students in particular, assessment criteria are important clues to determine the core content of their study programme (Sadler 1985). Students should therefore be provided with transparent assessment criteria before they tackle learning tasks. In many Dutch secondary vocational education programmes, students are required to select competencies they want to master from the above-mentioned list of 25 general competencies, which are also used as assessment criteria. It is questionable, however, whether these general, competency-based assessment criteria (e.g. 'informs the patient proactively') offer novice students a sufficiently sound basis for assessing their own performance on learning tasks, since these competencies are not tailored to the work context of their future profession. These questions regarding competency-based assessment criteria, coupled with the

fact that the nature of effective assessment criteria in competency-based education is an under-researched topic (Fastré et al. 2010), led to the study reported in this article. The main research question is: which assessment criteria are most effective in promoting student learning?

The study focused on two major differences between criteria: competency-based versus performance-based criteria, and presenting students with the relevant assessment criteria for the task at hand versus a list of all possibly relevant criteria. The questions arising from these differences in criteria are: Which type of assessment criteria lead to better performance and self-assessments? And do students perform better and are able to assess their own performance better when presenting them the relevant assessment criteria for the task at hand? A last research question is: What is the most optimal combination of relevance and type of criteria in order to stimulate better performance and self-assessments?

Competency-based and performance-based criteria

Within competency-based education, assessment criteria are often formulated as competencies; in other words, what the student is *able* to do (Grégoire 1997; Crossley and Jolly 2012). The language of these competencies has tended to adopt a prescriptive rather than a descriptive approach (Lurie 2012). An example is 'to be able to communicate properly', an ability that results from the integration of knowledge, skills and attitudes. Performance-based criteria, on the other hand, are formulated in terms of what the student *does*. An example is 'Explaining the goal to the patient in language the patient can understand', an action the student has to undertake. The difference between these two types of criteria should be seen as a continuum that is linked to different points in a training programme; in other words, they are related to different levels of training. Competency-based criteria are less meaningful to novice students than to advanced students, because novices have not yet achieved the requisite integration of knowledge, skills and attitudes (Fastré et al. 2010). There is, however, a direct relationship between competency-based and performance-based criteria, because the latter specify context-specific performance in relation to the competency in question (Crossley and Jolly 2012). For student nurses, for example, the performance-based criteria related to the competency 'communicating properly' are 'introduce yourself to the patient', 'tell the patient what you are going to do', etc. As students must be able to generalize what they have learned and be able to apply and extend their learning to a range of situations, it is important to teach them the link between performance and competencies early in their education. In this way, they can grow from thinking in more specific (performance-based) terms, to more abstract (competency-based) terms. This can be done by offering them a broad range of situations under varying circumstances in which they have to perform the same task. The technical knowledge for that task is constant, but the circumstances vary.

Figure 1 shows the continuum from performance-based to competency-based assessment criteria. Four dimensions determine the suitability of the different criteria for novice or more advanced students: one basic dimension runs from 'What *do* you do?' to 'What *can* you do?' The other three run (1) from behaviour that is directly observable to behaviour that requires interpretation in order to link it to performance, (2) from task-dependent descriptions to task-independent descriptions, and (3) from low investment of mental effort to high investment of mental effort.

Competency-based criteria address the student's ability to perform a certain task rather than his or her actual task performance. As a result, students first have to interpret the criteria in order to relate them to performance before they can make an accurate assessment

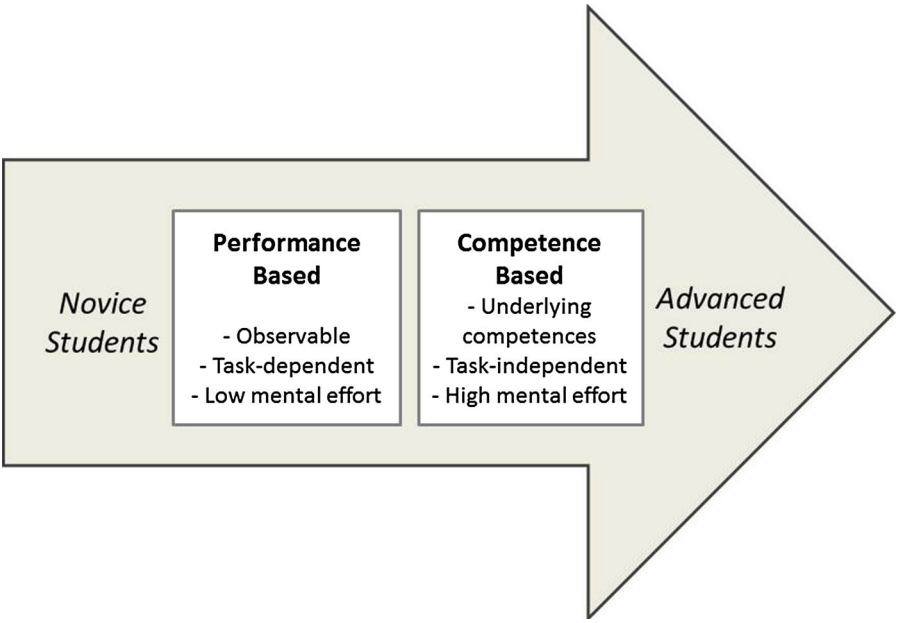

Fig. 1 Continuum of performance-based to competency-based assessment criteria

(Grégoire 1997). According to Lurie (2012), it is important to define assessment criteria in terms of the situations to which they are relevant, rather than as global personal characteristics. Performance-based criteria, on the other hand, pertain to behaviours that are readily observable. Task performance in accordance with these criteria provides some evidence that students have mastered the underlying competencies (Miller 1990). Performance-based criteria are expected to be more beneficial to novice students than competency-based criteria, because they do not require the prior knowledge that is indispensable for the correct interpretation of competency-based criteria.

The knowledge that is needed to link competency with performance can only be developed from experience with a series of different tasks. Furthermore, competencies go beyond concrete task performance and are best demonstrated through the acceptable performance of a variety of tasks. Thus, competency-based criteria are not dependent on a particular task (Albanese et al. 2010), unlike performance-based criteria, which are task-dependent. A different set of performance-based criteria may be relevant to each new task. Performance-based criteria are expected to be more beneficial to novice students, because of their task specificity and the concrete specification of what is expected of the student. This can also have positive effects on students' motivation, learning and performance (Ecclestone 2001).

Earlier research by Fastré et al. (2010) found that novice students invested relatively less mental effort in the assessment of their task performance when they were given performance-based criteria than when they were given competency-based assessment criteria. Because competency-based criteria are rather broadly defined compared to performance-based criteria, their interpretation requires students to exert additional mental effort to interpret and link them to the learning task at hand. Competency-based criteria therefore increase the cognitive load on the learner. Cognitive load theory presupposes that people have limited working memory capacity and it highlights the importance of avoiding 'extraneous' cognitive

load, that is, cognitive load that is not immediately relevant to learning (Sweller et al. 1998; Van Merriënboer and Sweller 2005). Performance-based criteria are less burdensome in terms of cognitive load, because they do not require the cumbersome process of interpretation before the actual assessment. This is particularly relevant to novice students, who have to expend considerable mental effort to link competency and performance because they have not yet acquired the necessary knowledge to translate competency into performance. Performance-based criteria reduce cognitive load, because they are directly observable and specify concretely what is expected from students (Fastré et al. 2010).

In summary, performance-based assessment criteria are expected to be more beneficial to novice students than competency-based criteria, because they are directly observable, clearly specify what is to be assessed, and require less mental effort.

Relevance of criteria

For complex learning tasks that are based on real-life tasks, there is a large set of potentially relevant assessment criteria, because not all tasks require the same behaviours (Sadler 1989). This set of criteria can be divided into two parts for each learning task: relevant criteria and irrelevant criteria. Students in competency-based vocational education are often given long lists of criteria, such as the 25 general competencies mentioned earlier, from which they have to select the criteria that are relevant to the task at hand (Kicken et al. 2008). This practice is based on the notion that for students to become independent learners, they must learn to make the distinction between relevant criteria and irrelevant criteria for the tasks they undertake.

Being confronted with a long list of all potentially relevant criteria, however, may have negative consequences for novice students, who are insufficiently equipped to identify which criteria are and which are not relevant to a specific task (Regehr and Eva 2006; Dunning et al. 2004). Unless the selection process is properly supported by a teacher or instructional materials, novice students are likely to randomly select some criteria and as a result assess their performance based on a mix of relevant and irrelevant criteria. It might be more effective to present novice learners with only the relevant criteria for the task at hand, because this allows them to focus their attention on comprehending and applying these criteria in order to arrive at an accurate assessment. This is in line with an earlier study by Fastré et al. (2012), in which novice students who were given only relevant assessment criteria reported investing more mental effort in assessments and showed higher task performance than students who had to make a selection from all potentially relevant criteria. This suggests that providing students with only relevant criteria can help them to focus on understanding and applying the criteria.

Hypotheses

Overall, it is hypothesized that presenting students with relevant performance-based assessment criteria will be most beneficial to their learning. It is also possible that specific combinations of criteria can contribute to significant learning gains.

The first hypothesis is that students who are given only relevant assessment criteria will show better task performance and better assessment skills than students who are given all potentially relevant assessment criteria. The second hypothesis is that students who are given performance-based criteria will show better task performance, better assessment skills and lower investment of mental effort than students who are given competency-based criteria. The third hypothesis is that a combination of relevant and performance-based criteria is most conducive to learning. We tested this by exploring the interaction effects of relevance and type of criteria.

Method

Participants

Ninety-three second-year students (7 males and 86 females) attending nursing and care programmes at three institutes of secondary vocational education participated in this study as part of their regular training in stoma care. In the Netherlands there are nursing education programs at both the level of vocational education and the level of higher education. At the vocational education level nurses are educated for performing less complex and routine-based tasks. These programs are at level 3 and 4 of the European Qualification Framework. Nursing education bachelor programs at the tertiary level, at level 6 in the European Qualification Framework, aim at educating nurses for specialized and more complex nursing tasks. Though there are differences, both types of programs pay attention to nursing skills such as stoma care.

The participating schools were selected on the basis of their interest in the research. We selected second-year students because they are relatively novice students in a 4-year program and because they are already somewhat familiar with performing nursing skills. All students attending the programs were selected, without further inclusion criteria. The mean age was 18.03 years ($SD = 1.01$) and the number of participants per institution was 32, 39 and 22, respectively. At each institute, participants were randomly assigned to one of four conditions: competency-based/all criteria ($n = 23$), competency-based/relevant criteria ($n = 23$), performance-based/all criteria ($n = 23$) and performance-based/relevant criteria ($n = 24$). The experiment was conducted in exactly the same way at all institutes. The task of stoma care was the next task in the program and was therefore selected. It is part of the core skills of the nursing task as a whole, but we could have selected any other nursing skill. The teachers of the students were involved in every step of the experiment and afterwards. One of the authors of this article is one of the teachers involved. Students could not withdraw from the experiment as it was part of their regular curriculum.

Materials

Figure 2 summarizes the materials described below.

Lecture

A 90-min lecture on the theoretical background to stoma care was developed. In order to ensure that the same content was taught at each institute, the materials were developed together with the teachers involved, and the same PowerPoint presentation was used for all lectures. The researchers examined the content and the delivery of the lectures given by the teachers and observed no differences between them.

Video examples and video assessment

An electronic learning environment was developed that contained six video fragments of around three minutes each, in which an expert nurse demonstrated a worked example of good stoma care. Studying a worked example is an effective way for novice learners to learn a new task (Stark et al. 2009). The six fragments presented the consecutive components of the whole task of stoma care: (1) introduction, (2) preparation, (3) removing the

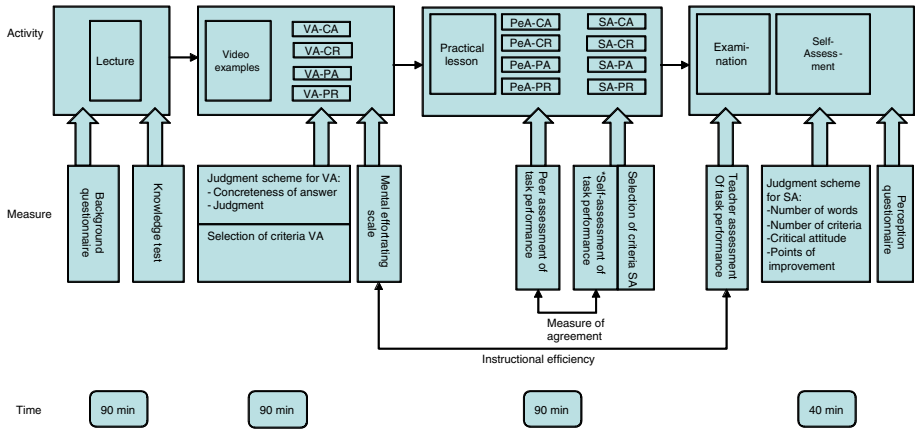

Fig. 2 Overview of materials and measures. *VA* video assessment, *PeA* peer assessment, *SA* self-assessment, *PA* performance-based/all criteria, *PR* performance-based/relevant criteria, *CA* competency-based/all criteria, *CR* competency-based/relevant criteria

old stoma bag, (4) attaching the new stoma bag, (5) finishing off care, and (6) evaluation and reporting. The students individually watched the video fragments on a computer screen. They were not allowed to stop the video, and they could watch each fragment a maximum of three times. On average, students watched the video fragments 1.19 times ($SD = .22$); an ANOVA revealed no significant differences between the conditions in this regard.

In order to help students make the assessment criteria more concrete, they were asked to describe the nurse's behaviour after watching each video fragment. They were given an electronic list of assessment criteria and asked to type a description of the nurse's concrete behaviours in relation to the relevant criteria in corresponding text boxes.

The list of assessment criteria was different for each of the four conditions, and the four groups of students used the criteria that were relevant to their condition during all parts of the study. In the competency-based/all-criteria condition, the assessment criteria were formulated as 22 stoma care competencies (VA–CA). The students were familiar with this type of criteria list, because it is routinely used during their training. Figure 3 shows examples of the criteria. The students were given a list of all potentially relevant criteria and were instructed to select the criteria they considered relevant to assessing the video fragments.

In the competency-based/relevant criteria condition, the students were given the same list of criteria, but the relevant criteria were highlighted and the students were asked to describe the nurse's behaviour in relation to only these criteria (VA–CR).

In the performance-based/all-criteria condition, the assessment criteria were formulated as 43 context-specific stoma care skills (VA–PA). Examples are given in Fig. 4. The students were given a list of all potentially relevant criteria and were instructed to select the criteria they considered relevant to assessing the video fragment.

The students in the performance-based/relevant criteria condition were given the same list of assessment criteria but the relevant criteria were highlighted, and the students were asked to describe the nurse's behaviour in relation to only these criteria (VA–PR).

The students had 90 min to complete the video assessment and spent on average 48.65 min ($SD = 11.96$) working in the electronic learning environment. An ANOVA

Criterion		How does the nurse show this criterion
Paying attention and sympathizing	Shows interest, listens actively, shows empathy to the patient	
	Puts herself in position of patients, colleagues and supervisors	
Working together and consulting	Informs the patient proactively	
	Consults patients and involves others on a regular basis and informs them	
	Makes agreements with patients and others involved in the task division	
Ethical and honest treatment	Acts in accordance with own norms, professional group, organization and legal constraints	
	Is honest, fair and acts without prejudice	
	Is discrete with sensitive topics	
	Communicates openly and clearly about sensitive topics	

Fig. 3 Screen dump of competency-based assessment criteria The criteria combined with a text box are the relevant criteria for a particular video fragment

Criterion		How does the nurse show this criterion
Prepares the patient	Introduces herself in an appropriate way	
	Explains the goal to the patient in language the patient can understand	
Consults the care file	Consults the care file for details about the stoma	
	Consults the patient for details about the patient	
Prepares the environment for the care	Takes action to ensure there is sufficient privacy	
	Collects the right materials	

Fig. 4 Screen dump of performance-based assessment criteria The criteria combined with a text box are the relevant criteria for a particular video fragment

revealed a significant difference between the conditions, with assessment time varying from 56.35 min ($SD = 17.41$) for the competency-based/all-criteria condition, 46.39 min ($SD = 9.80$) for the competency-based/relevant criteria condition, 46.74 min ($SD = 6.70$) for the performance-based/all-criteria condition, to 45.25 min ($SD = 8.18$) for the performance-based/relevant criteria condition. We did not use time on task as a covariate, because the students in the performance-based/relevant criteria condition, for which the most positive effects were predicted, spent the least amount of time on the learning task, which made the experiment more conservative.

Practice session, peer assessment and self-assessment

A 90-min practice session was developed, in which students practised the stoma care task in groups of two or three students, with a fellow student acting as the patient. After the practice session, the students scored their peers' task performance. The assessment was identical to that of the video fragments, except that the students were not asked to give a description of the behaviour (as was done for the video fragments) but to score their peers' performance in relation to the relevant criteria using a 4-point scale: (1) behaviour not shown, (2) behaviour shown but unsatisfactory, (3) behaviour shown and satisfactory, and (4) behaviour shown and good. The students were also asked to self-assess their own task performance in this way. Students in the all-criteria condition were asked to indicate which criteria they considered relevant to the task at hand.

Test and self-assessment

A test task was developed together with the teachers in which the students individually performed the task of stoma care with a simulated patient, that is, an actor who played the patient role. Throughout the course of their education, students initially practise with simulated patients, before practising with real patients. Simulated patients are often used in tests. Test task performance was videotaped to enable subsequent assessment by teachers. After completing the stoma care task, the students were given a blank form on which they had to assess their own task performance and indicate what went well and what went wrong.

Measures

Background questionnaire

Demographic data (age, sex and prior education) were collected using a short questionnaire. Students' perceptions of the relevance of the self-assessment and of their ability to self-assess were measured using the Self-Directed Learning Skills Questionnaire adapted from Kicken et al. (2006) and with proven reliability (Fastré et al. 2010). Table 1 shows the Cronbach's alpha scores of the perception scales. Internal consistencies were acceptable to high, ranging from .69 to .88. There were no significant demographic differences between the groups in the four conditions.

Knowledge test

After the lecture, a 15-item multiple choice test was administered to assess the students' knowledge of stoma care. The knowledge test was jointly set by the teachers to ensure its validity. An example item was: What are the three types of stomas? Possible responses were (a) calastoma, ilastoma, urastoma, (b) colostoma, ileostoma, urostoma, (c) colostoma, uleostoma, uristoma, and (d) culostoma, uleostoma, uristoma. On average, students scored 8.0 on a score of 10 on the knowledge test, with a standard deviation of 1.42. There were no significant differences between the groups, indicating that all students had the same level of knowledge at the end of the lecture and before the experimental manipulations.

Judgment scheme for the video assessment

The accuracy of the video assessments was judged by two raters (interrater reliability was acceptable at $r = .65$, $p = .00$) using a judgment scheme to specify the quality of the assessments. The overall score for the quality of the video assessments was the sum of the *z*-scores for each relevant criterion on two aspects: whether the student gave a concrete description of the nurse's behaviour (e.g. 'She reassures the patient by saying that it's difficult for the patient to have someone else do this') (0 = no concrete answer, 1 = concrete answer) and whether the student gave a judgment of the nurse's behaviour (e.g. 'She did very well because she constantly reassured the patient during the care process') (0 = no judgment, 1 = judgment). The higher the sum of the z-scores, the higher the score for the quality of the video assessment.

Selection of criteria during the video assessment

It was measured how successful the students in the all-criteria groups were in selecting criteria for the video assessments. The aim was for the students to select as many relevant and as few irrelevant criteria as possible.

Mental effort rating scale

After the assessments of the video fragments (six assessments in total), the students were asked to indicate the mental effort involved by rating the 'effort required to perform the assessment' on a 7-point scale as used in an experiment in secondary vocational education by Corbalan et al. (2009). The scale points ranged from very low mental effort (1) to very high mental effort (7). Corbalan et al. (2009) report Cronbach's alphas for the mental effort rating scale of between .94 and .96 and its validity is satisfactory (Paas et al. 2003). The mean of the six scores was calculated as the mental effort invested in assessing the video fragments.

Peer assessment of task performance

The students assessed their peers' task performance during the practice session using the 4-point scale from the practice session. Peer-assessed task performance was the mean score on the relevant assessment criteria.

Self-assessment of task performance

The self-assessment procedure was identical to that of the peer assessment. Self-assessed task performance was the mean self-assessment score for the relevant assessment criteria.

Selection of relevant criteria during self-assessment

For the students in the two all-criteria conditions, the number of relevant criteria and the number of irrelevant criteria selected during the self-assessment indicated their ability to select appropriate criteria when assessing their own performance.

Table 1 Reliability of the self-directed learning skills questionnaire

Scale	Cronbach's alpha	# items	Example item[a]
Relevance of self-assessment	.69	3	I think the opinion of the teacher is more important than self-assessment
Ability to self-assess	.88	12	I can assess to what extent my performance meets the assessment criteria

[a] Items have been translated from the Dutch

Teacher assessment of test task performance

Twenty teachers observed and assessed the students' videotaped performances on the stoma care test using the list of relevant performance-based criteria. All the teachers had attended a 4-h training session on performance assessment 1 week before the actual assessment in order to enhance interrater agreement. In this training, no data were used that were part of the data corpus but only example videos that were specifically prepared for the training program. The training consisted of five phases. In the first phase (30 min), the assessment instrument with the relevant performance-based criteria was described for the participants in detail, by explaining each criterion separately. In the second phase (45 min), three video-recordings of students who showed exemplary behaviour were shown and discussed; the observed behaviour was linked to each criterion and it was explained why this criterion was sufficiently met. Note that the videotaped performances were not segmented: The participants observed the whole video recording and assessed performance of the student using the list with relevant performance-based criteria, just as they would do in a real-life assessment. In the third phase (45 min), participants had to play out behaviours that did *not* meet the criteria and the assessment of these behaviours was discussed in the group. After a 15-min break, in the fourth and most important phase (90 min), three video-recordings of students who showed behaviours that did not meet all criteria were shown. Participants had to assess these students with the assessment instrument; for each of the three examples, their assessments were compared criterion by criterion and, in case of disagreements, discussed in the group until agreement was reached. Thus, as a result of the discussion, participants had to sharpen their interpretation of the criteria. In the fifth and final phase (15 min), questions were answered and participants received instructions for rating the video-recordings that resulted from the experiment (for a description of the training, also see (Fastré et al. 2010).

After data collection, video-recordings of students were assessed by 61 different pairs of teachers, each pair randomly selected from the 20 teachers who participated in the assessor training. Each pair assessed one or two video-recordings. This approach was taken because it reduced the time commitment for individual raters and thus offered the opportunity to have each student assessed by two assessors. We used Pearson correlations for computing interrater reliability and interpret $r < .50$ as unacceptable, $.50 < r < .70$ as acceptable, and $r > .70$ as good. The average correlation between the scores of all 61 pairs of teachers was $r = .53$ (ranging from .32 to .75; $SD = .11$). This shows that some correlations are rather low. Unfortunately, raters (i.e. teachers) were no longer available after completion of the experiment to recode the low-agreement cases.

Judgment scheme for self-assessment

The overall score for the quality of self-assessment of test task performance was the sum of the *z*-scores on the following aspects: the number of words used (word count), the number of relevant criteria selected (number of relevant criteria), the student's critical attitude towards his or her own performance (0 = no critical attitude, 1 = critical attitude) and the presence of points for improvement (0 = no points for improvement, 1 = points for improvement). The higher the sum of the *z*-scores, the higher the score on the quality of self-assessment. The quality of the self-assessments was judged by two raters. Interrater reliability was $r = .86$, $p = .00$.

Perception questionnaire

To ensure the fact that the effects we find in this study are a result of the differences in experimental conditions and not of differences in the perception of the learning environment, we measured student perceptions. The students evaluated their learning experience by rating the following aspects of perception of the learning environment on a 4-point Likert scale: interesting course material, task orientation, general pleasure and interest, pleasure and interest in relation to reflection, and usefulness. Higher scores indicate a more positive perception of the learning experience. Two scales (interesting course material and task orientation) were taken from the Inventory of Perceived Study Environment (IPSE; Wierstra et al. 1999). Three scales (general interest and pleasure, interest and pleasure in relation to reflection, and usefulness) were taken from the Intrinsic Motivation Inventory of Deci et al. (1994), translated into Dutch by Martens and Kirschner (2004). Table 2 shows the Cronbach's alpha scores of the perception scales; internal consistencies ranged from .67 to .89, which is considered acceptable to high.

Measure of agreement

Agreement between peer and self-assessment during the practice session was determined by computing the Pearson correlation coefficient.

Instructional efficiency

Instructional efficiency was determined by the relationship between test task performance and the average mental effort during the video assessments (Paas and van Merriënboer 1993; van Gog and Paas 2008). A lower mental effort (extraneous load) during the video assessments (at the time of training) has a positive effect on learning (van Merriënboer and Sweller 2005), which leads to a higher performance in the test task (at a later time). Performance and mental effort scores were standardized, and the *z*-scores were entered into the formula:

$$E = \frac{Z_{Performance} - Z_{Mental\,Effort}}{\sqrt{2}}$$

A combination of relatively low mental effort with relatively high test task performance was indicative of high instructional efficiency, whereas a combination of relatively high mental effort with relatively low test task performance was indicative of low efficiency.

Procedure

The demographics questionnaire was administered before the lecture and the multiple-choice knowledge test after the lecture.

Immediately after the lecture, the students were randomly assigned to a group corresponding to one of the four conditions; they remained in the same group for the duration of the study. The groups first assessed the video fragments (maximum time: 90 min) and immediately after that took part in the practice session followed by peer and self-assessments, which together took another 90 min.

One week after the practice session, the students took the test and assessed their own performance (40 min). As all groups had this same time frame, we have no reasons to assume that any external influences may have caused differences between groups for this matter. Student test task performance was subsequently assessed by two teachers. The evaluation questionnaire was administered immediately after the test.

Data analysis and variables

A two-way ANOVA was conducted to test for effects of relevance of criteria (only relevant vs. all potentially relevant criteria), type of criteria (performance-based vs. competency-based) and their interaction on the following variables: test task performance, quality of self-assessment (number of words, number of criteria, critical attitude, points for improvement), instructional efficiency, quality of video assessment (concreteness, judgment), judgment of video assessment, mental effort during the learning phase, peer assessment of task performance, and self-assessment of task performance. Furthermore, a two-way ANOVA was conducted to test for effects on the evaluation questionnaire variables: interesting course material, task orientation, interest and pleasure, interest and pleasure in reflection, and usefulness. For all analyses, the significance level was set to .05. Partial Eta-squared is given as a measure of effect size, with $\eta_p^2 = .01$ indicating a small effect, $\eta_p^2 = .06$ a medium effect and $\eta_p^2 = .14$ a large effect.

A t test was conducted for the selection of relevant and irrelevant criteria during the video assessments. Cohen's d was used as a measure of effect size, with $d \geq 1.3$ indicating a very large effect, $.80 > d > .29$ a large effect, $.50 > d > .79$ a medium effect, $.20 > d > .49$ a small effect, $-.19 > d > .19$ no effect, $-.20 > d > -.49$ a small negative effect, etc. They were only analysed in the all-criteria groups, because the other groups did not select criteria.

Results

The results on the dependent variables in the test phase and the learning phase, and the students' perceptions are reported consecutively.

Test phase

Table 3 presents the means and standard deviations for the dependent variables in the test phase.

Concerning the first hypothesis, a main effect of relevance ($F(1, 89) = 3.178$, $MSE = .022$, $p = .04$, $\eta_p^2 = .034$) was found for test task performance as assessed by the

Table 2 Reliability of the perception measures

Scale	Cronbach's alpha	# items	Example item[a]
Inventory of perceived study environment			
Interesting course materials	.67	7	The learning tasks are interesting
Task orientation	.68	3	I know what is expected of me when performing a task
Intrinsic motivation inventory			
Interest and pleasure in learning tasks	.70	7	I enjoy working on the learning tasks
Interest and pleasure in reflection	.89	9	I find it interesting to reflect
Usefulness	.72	4	I should like to conduct more learning tasks because they are useful

[a] Items have been translated from the Dutch

teacher, indicating better test task performance of the relevant-criteria groups ($M = 2.84$, $SD = .28$) compared to the all-criteria groups ($M = 2.71$, $SD = .32$). There was neither a main effect of type of criteria nor an interaction effect for test task performance. No significant main effects or interaction effects were found for quality of self-assessment or the related more specific variables. Concerning the second hypothesis, a main effect of type of criteria ($F(1, 89) = 9.483$, $MSE = 8.000$, $p = .00$, $\eta_p^2 = .096$) was found for instructional efficiency, indicating higher efficiency for the performance-based groups ($M = .29$, $SD = 1.01$) than for the competency-based groups ($M = -.30$, $SD = .81$). A marginally significant effect of relevance was found ($F(1, 89) = 1.817$, $MSE = 1.533$, $p = .09$, $\eta_p^2 = .020$), indicating a marginally but significantly higher instructional efficiency for the relevant-criteria groups ($M = .13$, $SD = .86$) compared to the all-criteria groups ($M = -.13$, $SD = 1.04$). No interaction effect was found for instructional efficiency.

Learning phase

Table 4 presents the means and the standard deviations for the dependent variables in the learning phase.

Linked to the first hypothesis, a main effect of relevance was found for the overall score on quality of the video assessments ($F(1, 89) = 12.517$, $MSE = 26.265$, $p = .00$, $\eta_p^2 = .123$). The relevant-criteria groups had a higher score on the quality of video assessment ($M = .53$, $SD = 1.54$) than the all-criteria groups ($M = -.54$, $SD = 1.38$).

Linked to the second hypothesis, a main effect of type of criteria was also found ($F(1, 89) = 3.632$, $MSE = 7.622$, $p = .03$, $\eta_p^2 = .039$), indicating a higher quality of video assessment in the performance-based groups ($M = .29$, $SD = 1.59$) compared to the competency-based groups ($M = -.29$, $SD = 1.47$). No interaction effect was found.

More specifically, a main effect of relevance was found ($F(1, 89) = 1.653$, $MSE = .890$, $p = .01$, $\eta_p^2 = .055$) for concreteness of answers in the video assessments, indicating that the relevant-criteria groups provided more concrete answers ($M = .85$, $SD = .36$) compared to the all-criteria groups ($M = .65$, $SD = .48$). In addition, a main effect of type of criteria was found ($F(1, 89) = 3.130$, $MSE = .538$, $p = .04$, $\eta_p^2 = .034$), indicating more concrete answers from the performance-based groups ($M = .83$, $SD = .38$) compared to the competency-based groups ($M = .67$, $SD = .47$).

Concerning the third hypothesis, a significant interaction effect for concrete answers shows that the positive effects of performance-based and relevant criteria occur primarily in the performance-based/relevant-criteria group ($F(1, 89) = 3.130$, $MSE = .538$, $p = .04$, $\eta_p^2 = .034$). A visual inspection of the interaction (Fig. 5) shows that the performance-based/relevant-criteria group gave more concrete answers than the other groups, all of which scored at roughly the same level.

A main effect of relevance was found for judgment of video assessment ($F(1, 88) = 3.908$, $MSE = .775$, $p = .03$, $\eta_p^2 = .043$), indicating that the relevant-criteria groups ($M = .36$, $SD = .49$) did and the other groups did not give a judgment of the nurse's behaviour ($M = .17$, $SD = .38$). There was no main effect of type of criteria and no interaction effect for judgment.

Concerning the second hypothesis, the selection of criteria during the video assessments was analysed only in the all-criteria groups, because the other groups did not select criteria. A *t*-test revealed a significant difference for the selection of relevant criteria ($t(44) = 17.68$, $p = .00$, $d = 5{,}22$), indicating that the performance-based group selected a much larger number of relevant criteria ($M = 34.61$, $SD = 3.74$) compared to the competency-based group ($M = 9.91$, $SD = 5.56$). A significant difference was also found for the selection of irrelevant criteria ($t(44) = -2.876$, $p = .01$, $d = -0{,}58$), with a much lower number of irrelevant criteria selected by the performance-based group ($M = 15.09$, $SD = 11.75$) compared to the competency-based group ($M = 49.39$, $SD = 55.98$).

Furthermore, a main effect of type of criteria was found ($F(1, 89) = 9.855$, $MSE = 8.917$, $p = .00$, $\eta_p^2 = .100$) for mental effort, indicating higher mental effort for the competency-based groups ($M = 3.61$, $SD = .95$) than for the performance-based groups ($M = 2.99$, $SD = .93$). There was no main effect of relevance and no interaction effect for mental effort.

The analysis of peer assessment of performance revealed a main effect of type of criteria ($F(1, 86) = 2.753$, $MSE = .603$, $p = .05$, $\eta_p^2 = .031$), indicating a higher score in the performance-based groups ($M = 3.45$, $SD = .53$) compared to the competency-based groups ($M = 3.29$, $SD = .38$). There was no main effect of relevance and no interaction effect.

Self-assessment of performance during the practice session showed a main effect of type of criteria ($F(1, 85) = 3.221$, $MSE = .953$, $p = .04$, $\eta_p^2 = .037$), indicating a higher score in the performance-based groups ($M = 3.33$, $SD = .53$) compared to the competency-based groups ($M = 3.12$, $SD = .56$). A moderate agreement between peer assessment and self-assessment was found, $r = .50$, $p = .00$, indicating congruity of self-assessments and peer assessments.

The selection of relevant and irrelevant criteria during self-assessment in the practice session was analysed only in the all-criteria groups. A *t*-test revealed a significant difference for relevant criteria ($t(42) = 3.793$, $p = .00$, $d = 1{,}17$), indicating that students in the performance-based group selected more relevant criteria ($M = 31.50$, $SD = 14.16$) compared to the competency-based group ($M = 19.36$, $SD = 4.78$). No significant difference was found for irrelevant criteria.

Evaluation questionnaire

Overall, students perceived the learning environment as interesting and useful. Table 5 shows the means and standard deviations for all scales.

On task orientation, a marginal effect of relevance was found ($F(1, 85) = 3.031$, $MSE = .831$, $p = .09$, $\eta_p^2 = .034$), with a higher perceived task orientation in the all-

criteria groups ($M = 3.36$, $SD = .51$) compared to the relevant-criteria groups ($M = 3.16$, $SD = .55$). A marginal interaction effect for task orientation, however, shows that the positive effect of providing all criteria is limited to the competency-based groups ($F(1, 85) = 2.967$, $MSE = .813$, $p = .09$, $\eta_p^2 = .034$). Visual inspection of the interaction effect (Fig. 6) reveals that the competency-based/all-criteria group shows higher perceived task orientation than the other three groups, which have roughly similar scores. Thus, only students in the competency-based/all-criteria group had a higher task orientation. No main or interaction effects were found for interesting course material, general interest and pleasure, interest and pleasure in relation to reflection, and usefulness.

Discussion

This present study investigated one of the problems inherent to the assessment of competencies, which is in today's nursing programs a frequently discussed subject. It contributes to the field of nursing in at least three ways. First, it shows how well-designed assessment instruments can encourage students' active involvement in assessments (Garside et al. 2009), with the emphasis on the value of self-assessment and peer assessment (Dearnley and Meddings 2007). Second, it explores how formative assessments can foster learning (Koh 2008) and, in particular, how they can strengthen the value of assessment for students' future learning experiences and lifelong learning (Leung et al. 2008). Third, it exemplified the advantages of high-fidelity simulation for assessment purposes (Alinier et al. 2006).

The general goal of this study was to investigate the effects of relevance and type of assessment criteria in relation to students' test task performance and assessment skills. It will first be discussed whether students perform better and are able to assess themselves better when presenting them the relevant assessment criteria for the task at hand. The first hypothesis—stating that students who are given only the relevant criteria will show higher task performance and make better assessments than students who are given all criteria—is largely confirmed by the results. As expected, relevant criteria were associated with higher test task performance. Students who were given relevant criteria seemed to know better what was expected of them, as indicated by the higher quality of their video assessments and more frequent use of judgments, although they did not show better self-assessment of test task performance. Relevant criteria were also associated with higher instructional efficiency.

The positive relationship between relevant criteria and test task performance confirms the results of an earlier study by Fastré et al. (2012). The absence of an effect of relevance of criteria on mental effort suggests that cognitive load may be of a different nature depending on the relevance of criteria (Sweller et al. 1998). For the all-criteria groups, the invested mental effort may have been 'extraneous', that is, ineffective for learning, with students engaging in unsuccessful efforts to distinguish between relevant and irrelevant criteria. In contrast, the invested mental effort may have been effective, viz. 'germane', for the relevant criteria groups, who managed to grasp and apply the criteria, as is evidenced by their higher test task performance and instructional efficiency.

The second discussion point regards the question which type of assessment criteria leads to better performance and self-assessments. The second hypothesis—stating that students who are given performance-based criteria will show higher task performance, make better assessments and invest less mental effort in the assessment than students who are given competency-based criteria—is also largely confirmed by the results. Students who were

Table 3 Means and standard deviations for dependent variables in the test phase

	Competency-based/all criteria ($n = 23$)		Competency-based/relevant criteria ($n = 23$)		Performance-based/all criteria ($n = 23$)		Performance-based/relevant criteria ($n = 24$)	
	M	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
Test task performance	2.67	.19	2.81	.28	2.76	.41	2.85	.29
Quality of self-assessment	−.31	3.05	.45	3.52	−.38	3.17	−.38	2.66
Number of words	86	48	87	53	73	45	74	36
Number of criteria	5.13	2.65	5.39	5.31	5.17	4.75	4.52	2.73
Critical attitude	.91	.29	.96	.21	.87	.34	.96	.21
Points of improvement	.22	.42	.17	.39	.09	.29	.22	.42
Instructional efficiency	−.46	.76	−.13	.85	.20	1.19	.38	.82

Table 4 Means and standard deviations for dependent variables in the learning phase

	Competency-based/all criteria ($n = 23$)		Competency-based/relevant criteria ($n = 23$)		Performance-based/all criteria ($n = 23$)		Performance-based/relevant criteria ($n = 24$)	
	M	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
Quality of video assessment	−.84	1.31	.25	1.44	−.24	1.41	.79	1.61
Concreteness of answer	.65	.49	.70	.47	.65	.49	1.00	.00
Judgment	.17	.39	.35	.49	.17	.39	.38	.49
Number of relevant criteria video assessment	9.91	5.56	–	–	34.61	3.74	–	–
Number of irrelevant criteria video assessment	49.39	55.97	–	–	15.09	11.75	–	–
Average mental effort during learning phase	3.61	1.08	3.61	.84	2.99	.93	2.99	.95
Task performance scored by peer	3.26	.42	3.32	.34	3.44	.58	3.47	.50
Task performance scored by self	3.02	.56	3.23	.52	3.40	.56	3.26	.51
Number of relevant criteria self-assessment	19.36	4.78	–	–	31.50	14.16	–	–
Number of irrelevant criteria self-assessment	17.59	6.87	–	–	17.05	14.07	–	–

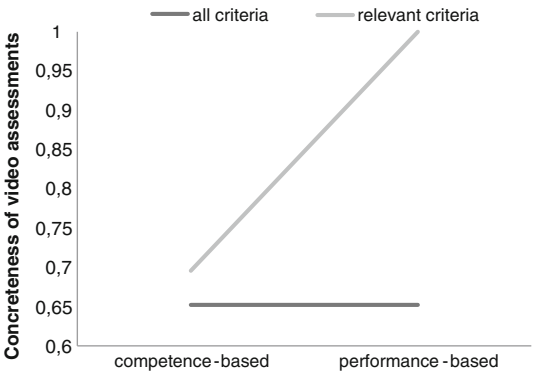

Fig. 5 Difference in concreteness of video assessments

given performance-based assessment criteria did indeed show higher task performance during the practice session as assessed by themselves and their peers. However, during the test task, performance as assessed by the teachers did not show this difference. Compared to the students in the competency-based criteria groups, the students in the performance-based criteria groups provided higher-quality video assessments, selected more relevant criteria and fewer irrelevant criteria during the video assessments, and selected more relevant criteria during the self-assessments in the practice session. Furthermore, instructional efficiency was superior for the performance-based criteria groups, which invested less mental effort in the video assessments but achieved a comparable test task performance.

The fact that performance-based criteria resulted in better assessments may be explained by the directly observable and task-specific character of these criteria (Gulikers et al. 2008). This confirms the results of Lurie (2012), who indicates that it is important to define assessment criteria in terms of the situations to which they are relevant, rather than as global personal characteristics. The students who used performance-based criteria performed better in the practice session, during which they could consult the list of performance-based criteria, providing guidance as to what was important in the assessment. One could argue that it is not really surprising that students with the performance-based version of the criteria performed better during practice than those who were given the competency-based version. Quite similar thoughts were already reported decades ago (see, e.g., the work of Mager 1975 and Duchastel and Merrill 1973). The findings of the present study are in line with these previous publications and also add to our understanding of the kinds of concrete materials and instructions that students need in order to fully realize the potential benefits of performance-based criteria.

The absence of a similar difference in test task performance and quality of assessment may be attributable to the students not having access to the criteria lists during the test. The time given to the students in this study to practise the assessment with recourse to the criteria list may have been too short for them to retain the criteria and apply them to improve test task performance. In earlier research (Fastré et al. 2010), the provision of performance-based criteria actually improved test task performance. Although this finding was not replicated in the present study, it is in line with the lower mental effort expended by the groups using performance-based criteria. Additionally, performance-based criteria were associated with higher instructional efficiency because of the favourable ratio of mental effort and test task performance (van Gog and Paas 2008).

The third discussion point regards the last research question, stating: What is the most optimal combination of relevance and type of criteria in order to stimulate better performance and self-assessments? The third hypothesis—that a combination of relevant and performance-based criteria is most conducive to learning—is partly confirmed by the data. An interaction effect of relevance of criteria and type of criteria was found for the concreteness of the video assessments. In line with our expectations, the performance-based/relevant-criteria group gave more concrete answers than the other three groups. The only other interaction that was found pertains to the evaluation of the learning environment, and in particular to task orientation with higher scores of the competency-based/all-criteria group. This is probably a result of the strong similarity of the criteria given to this group with the assessment instrument they were familiar with from their regular educational programme, such as the list of 25 general competencies. For the other groups, the assessment criteria were unfamiliar. There were no other differences in perceptions of the learning environment, which were positive for all the groups.

Future research should examine the relationship between the relevance of criteria and the type of cognitive load (germane or extraneous) experienced by students. As a related issue, the effects of the relevance of criteria should be examined among more advanced students. As students progress through a programme, they are likely to become better able to translate competencies into desired performances, which will reduce the cognitive load associated with competency-based criteria. As students progress, they should learn to concentrate on the performance as a whole rather than as isolated tasks to be done, which pleads for more competency-based criteria. Additionally, the cognitive load in having to deal with a list of all possible criteria may become more germane for more advanced students, and contribute to their ability to distinguish relevant from irrelevant criteria.

A shortcoming of this study is the limited duration of the intervention, which was restricted to one single learning task. The achievement of higher levels of the complex skill of self-assessment requires substantial training (van Merriënboer and Kirschner 2007). Although the intervention had a limited duration, an increase in the skill of self-assessment was observed. However, this increase was quite modest. It is expected that a longer duration of the study might have caused larger effects.

The self-assessment format, with one open question, may also have been suboptimal, because this type of question may have elicited short and superficial responses, making it difficult to find differences between the groups. More important, the assessment of the test tasks showed weak interrater reliability according to accepted ranges (Nunnally and Bernstein 1994). There might be at least two reasons for this. First, reaching agreement between a rather large group of assessors is difficult and more extensive training than the provided 4-h session may be required (Hoyt and Kerns 1999). Second, it may possibly be a result of the adopted approach. We choose for an approach were each student was assessed by two assessors. Although the average interrater reliability score was rather low but acceptable, the coding of the data by the pairs yet differed substantially. For future research, it is advisable to pay even more attention to the training of raters, to use more than two raters per student, and to conduct generalizability studies in order to heighten the reliability of test task assessments. Now, the findings of our study concerning the test tasks should be cautiously interpreted until these are reinforced by future studies.

This study was conducted among students from three institutes for secondary vocational education, which supports the generalisation of the findings to a broader perspective. Practical implications resulting from this study are that novice students should preferably receive relevant, performance-based assessment criteria in order to improve their assessment skills and task performance. In today's practice, where students in competency-based

Table 5 Means and standard deviations for evaluation questionnaire

	Competency-based/all criteria ($n = 23$)		Competency-based/relevant criteria ($n = 22$)		Performance-based/all criteria ($n = 22$)		Performance-based/relevant criteria ($n = 24$)	
	M	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
Interesting course material	3.39	.29	3.27	.44	3.30	.46	3.35	.37
Task orientation	3.52	.47	3.13	.50	3.19	.50	3.19	.61
Interest and pleasure	3.72	.28	3.55	.44	3.57	.37	3.50	.41
Interest and pleasure in reflection	2.67	.74	2.79	.58	2.73	.58	2.60	.59
Usefulness	3.48	.40	3.39	.44	3.52	.45	3.44	.44

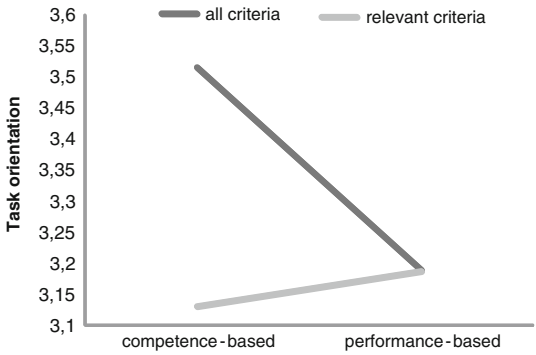

Fig. 6 Difference in perceived task orientation

vocational education in the Netherlands are often given long lists of criteria, such as the 25 general competencies, they have to select the criteria that are relevant to the task at hand themselves (Kicken et al. 2008). This may be a barrier to effective learning because it does not help them to learn what acceptable task performance looks like. Although this study was a relatively short intervention in the study program, a practical implication is that the rest of the program needs to be adjusted. However necessary, it should be noted that the process of formulating performance-based criteria and determining their relevance to each learning task is a time-consuming process. It is not very likely that without training and further support teachers are able to formulate and use the assessment criteria in a proper way. Our study does not imply that competency-based criteria are useless but they are not supportive for novice students. More advanced students must learn to connect competencies to performance and to distinguish between relevant and irrelevant competencies, because these skills are indispensable for their professional development. In the professional practice of nursing, the concept of competencies is used for appraisal and performance feedback purposes, so student nurses need to be prepared to think and reflect in terms of competencies. It should be noted, however, that the implementation of competency-based assessments in secondary vocational education is still in its infancy and faces many different problems. The added value of competency-based assessment has not been proven yet (see, for example, the studies of Biemans et al. 2009; Gulikers et al. 2010).

In that sense we argue that competency-based criteria are necessary in the perspective of lifelong learning and to create transfer to other contexts, but they are not the most suitable criteria for novice students because for them they create a high cognitive load. This leads us to the following implication concerning the need to change from performance-based towards competency-based criteria in a gradual manner. In the beginning students need considerable support by giving them the relevant and performance-based criteria; although these criteria may not be beneficial on the long term, they yield a manageable cognitive load and provide a stepping stone to the next phase. In this next phase, students have become skilled in the use of performance-based criteria and can gradually shift to selecting the relevant criteria themselves and to understand the underlying competencies; these criteria are relevant on the long term and prepare students for lifelong learning. Further research should provide indications on how to handle this transition.

Although this research was conducted in nursing education, a domain in which considerable attention is paid to assessment research as shown earlier, our results are not restricted to nursing education as such. We expect these results to be equally applicable to other domains in vocational education. There are two reasons for this. First, as indicated

before, all educational programs in Dutch senior vocational education use the same set of 25 competencies, which makes it relatively easy to adjust assessment instruments that have been developed in one domain to other domains (e.g., from the domain "welfare and care" to domains such as "tourism and recreation" or "trade and entrepreneurship"). Second, all domains in vocational education are characterized by professional skills that can be specified by performance-based criteria. Thus, a gradual transition from performance-based to competence-based assessment can be realized in all different domains in vocational education.

To conclude, the introduction of competency-based education and competency-based assessment forces educators to reconsider the assessment criteria used in their programmes. Our results show that novice students are not yet able to work with abstract competency-based criteria and select relevant criteria for specific tasks. It is far more beneficial to offer novice students relevant and performance-based criteria at the beginning of their study programme.

References

Albanese, M. A., Mejicano, G., Anderson, W. M., & Gruppen, L. (2010). Building a competency-based curriculum: The agony and the ecstasy. *Advances in Health Sciences Education, 15*(3), 439–454.

Alinier, G., Hunt, B., Gordon, R., & Harwood, C. (2006). Effectiveness of intermediate-fidelity simulation training technology in nursing education. *Journal of Advanced Nursing, 54*(3), 359–369.

Biemans, H., Nieuwenhuis, L., Poell, R., Mulder, M., & Wesselink, R. (2004). Competence-based VET in the Netherlands: Background and pitfalls. *Journal of Vocational Education and Training, 54*(4), 523–538.

Biemans, H., Wesselink, R., Gulikers, J., Schaafsma, S., Verstegen, J., & Mulder, M. (2009). Towards competence-based VET: Dealing with the pitfalls. *Journal of Vocational Education & Training, 61*(3), 267–286.

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

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

COLO. (2008). Prepared for the future: Dutch qualifications for the labour market. Retrieved from http://www.colo.nl/publications.html.

Corbalan, G., Kester, L., & van Merriënboer, J. J. G. (2009). Dynamic task selection: Effects of feedback and learner control on efficiency and motivation. *Learning and Instruction, 19*, 455–465.

Crossley, J., & Jolly, B. (2012). Making sense of work-based assessment: Ask the right questions, in the right way, about the right things, of the right people. *Medical Education, 46*(1), 28–37.

Dearnley, C. A., & Meddings, F. S. (2007). Student self-assessment and its impact on learning. A pilot study. *Nurse Education Today, 27*, 330–340.

Deci, E. L., Eghrari, H., Patrick, B. C., & Leone, D. (1994). Facilitating internalization: The self-determination theory perspective. *Journal of Personality, 62*, 119–142.

Duchastel, P. C., & Merrill, P. F. (1973). The effects of behavioural objectives on learning: A review of empirical studies. *Review of Educational Research, 43*, 53–69.

Dunning, D., Heath, C., & Suls, J. M. (2004). Flawed self-assessment: Implications for health, education, and the workplace. *Psychological Science in the Public Interest, 5*(3), 69–106.

Ecclestone, K. (2001). 'I know a 2:1 when I see it': Understanding criteria for degree classifications in franchised university programmes. *Journal of Further and Higher Education, 25*(3), 301–313.

Fastré, G. M., van der Klink, M., & van Merriënboer, J. J. G. (2010). The effects of performance-based assessment criteria on student performance and self-assessment skills. *Advances in Health Sciences Education, 15*(4), 517–532.

Fastré, G. M., van der Klink, M., van Merriënboer, J. J. G., & Sluijsmans, D. (2012). Drawing students' attention to relevant assessment criteria: Effects on self-assessment skills and performance. *Journal of vocational education and training, 64*(2), 185–198.

Garside, J., Nhemachena, J. Z. Z., Williams, J., & Topping, A. (2009). Repositioning assessment: Giving students the 'choice' of assessment methods. *Nurse Education in Practice, 9*, 141–148.

Grégoire, J. (1997). Diagnostic assessment of learning disabilities: From assessment of performance to assessment of competence. *European Journal of Psychological Assessment, 13*(1), 10–20.

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

Gulikers, J. T. M., Kester, L., Kirschner, P. A., & Bastiaens, Th J. (2008). The influence of practical experience on perceptions, study approach and learning outcomes in authentic assessment. *Learning and Instruction, 18*(2), 172–186.

Hoyt, W. T., & Kerns, M. D. (1999). Magnitude and moderators of bias in observer ratings: A meta analysis. *Psychological Methods, 4*, 403–424.

Kicken, W., Brand-Gruwel, S., & van Merriënboer, J. J. G. (2006). *Self-directed learning skills questionnaire*. Heerlen, the Netherlands: Open Universiteit Nederland.

Kicken, W., Brand-Gruwel, S., & van Merriënboer, J. J. G. (2008). Scaffolding advice on task selection: A safe path toward self-directed learning in on-demand education. *Journal of Vocational Education and Training, 60*, 223–239.

Koh, L. C. (2008). Refocusing formative feedback to enhance learning in pre-registration nurse education. *Nurse Education in Practice, 8*, 223–230.

Leung, S. F., Mok, E., & Wong, D. (2008). The impact of assessment methods on the learning of nursing students. *Nurse Education Today, 28*, 711–719.

Lurie, S. J. (2012). History and practice of competency-based assessment. *Medical Education, 46*, 49–57.

Mager, R. (1975). *Preparing instructional objectives* (2nd ed.). Belmont, CA: Lake Publishing Co.

Malone, K., & Supri, S. (2012). A critical time for medical education; the perils of competence-based reform of the curriculum. *Advances in Health Science Education, 17*, 241–246.

Martens, R. L. & Kirschner, P. A. (2004). What predicts intrinsic motivation? Paper presented at the 2004 AECT convention, Chicago.

Miller, G. E. (1990). The assessment of clinical skills/competence/performance. *Academic Medicine, 65*, S63–S67.

Norcini, J., Anderson, B., Bollela, V., Burch, V., Costa, M. J., Duvivier, R., et al. (2011). Criteria for good assessment: Consensus statement and recommendations from the Ottawa 2010 Conference. *Medical Teacher, 33*(3), 206–214.

Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric theory*. New York, NY: McGraw-Hill.

Paas, F., Tuovinen, J. E., Tabbers, H., & van Gerven, P. (2003). Cognitive load measurement as a means to advance cognitive load theory. *Educational Psychologist, 38*(1), 63–71.

Paas, F., & van Merriënboer, J. J. G. (1993). The efficiency of instructional conditions: An approach to combine mental effort and performance measures. *Human Factors, 35*, 737–743.

Regehr, G., & Eva, K. (2006). Self-assessment, self-direction, and the self-regulating professional. *Clinical Orthopaedics and Related Research, 449*, 34–38.

Sadler, D. R. (1985). The origins and functions of evaluative criteria. *Educational Theory, 35*(3), 285–297.

Sadler, D. R. (1989). Formative assessment and the design of instructional systems. *Instructional Science, 18*, 119–144.

Stark, R., Kopp, V., & Fischer, M. R. (2009). Case-based learning with worked examples in complex domains: Two experimental studies in undergraduate medical education. *Learning and Instruction*. doi:10.1016/j.learninstruc.2009.10.001.

Sweller, J., van Merriënboer, J. J. G., & Paas, F. (1998). Cognitive architecture and instructional design. *Educational Psychology Review, 10*, 251–294.

van der Klink, M. R., Boon, J., & Schlusmans, K. (2007). Competences and vocational higher education: Now and in future. *European Journal of Vocational Training, 30*(1), 67–82.

Van Gog, T., & Paas, F. (2008). Instructional efficiency: Revisiting the original construct in educational research. *Educational Psychologist, 43*, 16–26.

Van Merriënboer, J. J. G., & Kirschner, P. A. (2007). *Ten steps to complex learning*. Mahwah, NJ: Erlbaum/Taylor and Francis.

Van Merriënboer, J. J. G., & Sweller, J. (2005). Cognitive load theory and complex learning: Recent developments and future directions. *Educational Psychology Review, 17*, 147–177.

Wierstra, R. F. A., Kanselaar, G., van der Linden, J. L., & Lodewijks, H. G. L. C. (1999). Learning environment perceptions of European university students. *Learning Environments Research, 2*, 79–98.