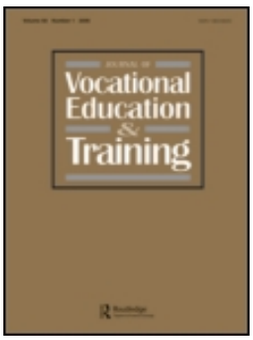

Routledge
Taylor & Francis Group

Journal of Vocational Education & Training

ISSN: 1363-6820 (Print) 1747-5090 (Online) Journal homepage: http://www.tandfonline.com/loi/rjve20

Drawing students' attention to relevant assessment criteria: effects on self-assessment skills and performance

Greet M.J. Fastré , Marcel R. van der Klink , Dominique Sluijsmans & Jeroen J.G. van Merriënboer

To cite this article: Greet M.J. Fastré , Marcel R. van der Klink , Dominique Sluijsmans & Jeroen J.G. van Merriënboer (2012) Drawing students' attention to relevant assessment criteria: effects on self-assessment skills and performance, Journal of Vocational Education & Training, 64:2, 185-198, DOI: 10.1080/13636820.2011.630537

To link to this article: https://doi.org/10.1080/13636820.2011.630537

Journal of Vocational Education and Training
Vol. 64, No. 2, June 2012, 185–198

Routledge
Taylor & Francis Group

Drawing students' attention to relevant assessment criteria: effects on self-assessment skills and performance

Greet M.J. Fastré[a,b,*], Marcel R. van der Klink[a], Dominique Sluijsmans[a,b,c], and Jeroen J.G. van Merriënboer[a,b]

[a]*Centre for Learning Sciences and Technologies, Open University of the Netherlands, Heerlen, The Netherlands;* [b]*Department of Educational Development and Research, Maastricht University, Maastricht, The Netherlands;* [c]*Faculty of Education, HAN University, Nijmegen, The Netherlands*

(*Received 15 November 2010; final version received 4 October 2011*)

We conducted a study among students in secondary vocational education programmes in nursing and care ($N = 68$). The students work on learning tasks, self-assess their task performance and formulate points for improvement. We compared two groups of students on self-assessment, identification of points of improvement and perceived effort for the assessment task. One group was given a list of all possible assessment criteria for all tasks in which the relevant criteria for a particular learning task were highlighted. The other group received the same list without highlighting. Students in the relevant criteria group outperformed the students in the all criteria group on a test task, but they experienced higher mental effort in self-assessing their performance. Care students in the relevant criteria group generated more points of improvement than care students in the all criteria group. Nursing students outperformed care students on the test task and care students selected more criteria than did nursing students.

Keywords: assessment; competence; standards; vocational education and training; curriculum innovation

Introduction

Nurses are expected to perform multiple tasks, which involve high cognitive load, and this often leads to stressful situations (Jourdain and Chênevert 2010). Apart from tasks in direct patient care, nurses have tasks in indirect patient care, such as keeping records, preparing medications and coordinating care (Wolf et al. 2006). In order to be able to adequately perform all these tasks, student nurses must acquire various professional competences for adequate performance in the workplace. They also have to develop self-directed learning skills to be able to self-assess their performance and judge their own strengths and weaknesses in order to cope with the uncertain, unpredictable and constantly changing circumstances of patient care now and in the future (Boud and Falchikov 2006; Kicken et al. 2009).

In this study, student self-assessment is defined as selecting relevant performance criteria from a predefined set of criteria, judging the extent to which performance of learning tasks meets relevant criteria, and identifying areas for improvement that are

*Corresponding author. Email: greetfastre@hotmail.com

ISSN 1363-6820 print/ISSN 1747-5090 online

http://dx.doi.org/10.1080/13636820.2011.630537
http://www.tandfonline.com

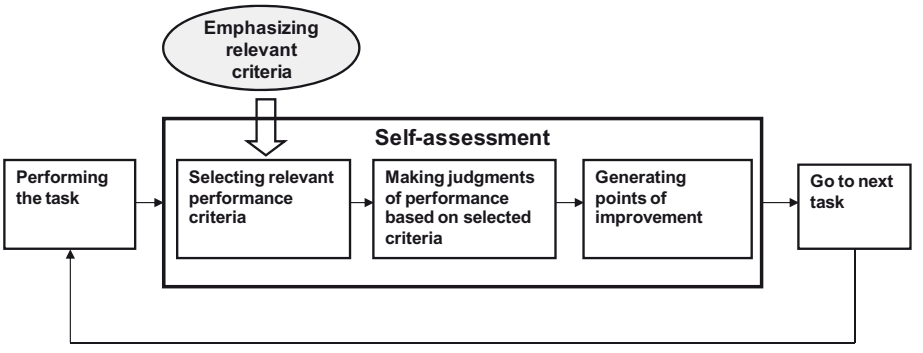

Figure 1. Model of the cyclical learning process with self-assessment at the heart.

to be addressed in future learning tasks. Figure 1 provides a schematic representation of the definition of self-assessment that is used in this study. It is based on work of Eva and Regehr (2005), and defines self-assessment as a cyclical process in which outcomes of self-assessments of previous tasks becomes input for subsequent tasks. Outcomes of self-assessment can be used as points of improvement to direct learners focus to particular aspects of a subsequent learning task.

According to Boud (1999), students' self-assessment is often based on cues provided by teachers and peers when students work on learning tasks. Self-assessment is therefore not an individual and isolated activity, but an integral part of a social learning process. The main question of the study reported in this article is: Does supporting students in selecting relevant performance criteria for assessing learning tasks help them to improve their task performance and self-assessment skills?

Quite a few studies have investigated students' skills in assessing their own performance (for reviews, see Boud and Falchikov 1989; Dochy, Segers, and Sluijsmans 1999; Dunning, Heath, and Suls 2004; Falchikov and Boud 1989). These studies show that the relationship between the actual quality and students' assessment of their performance is often rather weak. There are at least three explanations for this. Firstly, most students have little or no experience with self-assessment. In their prior education, nursing students have mostly been passive rather than active actors in assessment, and they have not been encouraged to take responsibility for the assessment process; they only took tests and received feedback on their performance (Boud and Falchikov 2006). Secondly, students' judgements are often biased. They have difficulty recognising their own incompetence (Kruger and Dunning 1999), and less experienced students usually overestimate their own performance (Dunning et al. 2004). Thirdly, students often rate their work differently compared to their teachers, and this may be increased by an absence of explicit performance criteria for self-assessment (Boud and Falchikov 1989).

In this study, students self-assessed their skills and received feedback on the quality of their self-assessment by comparing it with assessment by the teacher. The study described in this article investigated students' inability to determine which performance criteria are relevant for a particular task, in other words, the first step of the cyclical process of self-assessment (see Figure 1; see also Orsmond, Merry, and Reiling 2002). In addition, two sub-problems were addressed: students' inability to: (1) independently identify appropriate criteria; and (2) select appropriate criteria for a particular learning task from the full set of criteria for all tasks in the

domain. Skills for selecting criteria are important and need to be trained if students are to develop into self-directed learners and nurses.

With regard to the first sub-problem, Boud and Falchikov (1989) state that the ability to determine which performance criteria are relevant for judging one's own performance is an important and integral part of self-assessment. Learning this skill requires training and opportunities to practise (Boud and Brew 1995; Orsmond et al. 2002). Dunning et al. (2004) point out that it is very difficult for novices to formulate criteria for their own performance. They make mistakes due to inadequate domain expertise, which they are unable to recognise because their lack of expertise prevents them from having a clear view of what good performance should entail. Student nurses with no experience in patient care, for example, do not know which competences are important in patient contacts or what constitutes 'good practice' for these contacts. It is therefore appropriate to use pre-specified performance criteria for novice students.

As for the second sub-problem, there is a very large set of performance criteria that students can reasonably consider with regard to performance and assessment of an authentic whole task that is representative of professional life (Sadler 1989). For each task, this whole set can be split up into relevant and irrelevant criteria. Student nurses have to master all the competences of their future work, which means their performance has to meet a substantial number of criteria. However, washing a highly demented male patient involves a different subset of criteria compared to feeding a 20-year-old female patient. Sadler (1989) argues that it is important for students to become competent in determining which criteria are relevant and which criteria are irrelevant for a particular task. When students select assessment criteria, four different situations are possible: (1) they correctly select a relevant criterion and identify it as relevant (true positive); (2) they incorrectly select an irrelevant criterion and identify it as relevant (false positive); (3) they incorrectly select a relevant criterion and identify it as irrelevant (false negative); and (4) they correctly select an irrelevant criterion and identify it as irrelevant (true negative). Ideally, students should only select true positives and avoid false positives.

In Dutch vocational education, students are often given long lists of pre-specified criteria without receiving information or training to help them determine which criteria are relevant for a particular task (Kicken, Brand-Gruwel, and van Merriënboer 2008; Corbalan, Kester, and van Merriënboer 2006). Regehr and Eva (2006) and Dunning et al. (2004) describe the risk of overload when students have to select relevant criteria for a task from a very long list. Students are likely to select criteria relating to skills they can perform well or enjoy performing, because people naturally seek positive emotions and therefore have difficulty recognising inadequacies in their performance. Consequently, students need to be explicitly stimulated to learn to select relevant criteria for tasks at hand.

Learning to select relevant criteria is complicated by the human mind's limited processing capacity; because numerous combinations of relevant criteria can be selected from the whole set of possibly relevant criteria (van Merriënboer and Sweller 2005). Moreover, performing a nursing task involves high cognitive load (Wolf et al. 2006), the so-called intrinsic cognitive load (Sweller, van Merriënboer, and Paas 1998). Selecting relevant criteria for assessing a specific task may require so much mental effort that students' cognitive capacity does not suffice to conduct the task and assess their task performance using those criteria. Moreover, selecting relevant criteria can be a load on students' working memory which can be either

positive or negative (Sweller et al. 1998). Positive load is caused by learning that leads to better comprehension of criteria (germane load), better task performance and more accurate self-assessment, such as constructing cognitive schemas linking particular features of tasks to the relevance or irrelevance of specific criteria. When student nurses are given the whole list of possibly relevant criteria they may be aware that a combination of technical nursing skills and communication skills is important, and thus apply appropriate criteria for communication skills also when giving a patient an injection. The load on working memory is negative when it is caused by processes that are not conducive to learning (extraneous load), such as when students incorrectly identify irrelevant criteria as relevant. For example, when students see the whole list of criteria they may be tempted to select irrelevant criteria simply because they are listed, and thus incorrectly apply criteria for communication skills (e.g. talking) when they are not applicable (e.g. a deaf patient). Whether the load is positive or negative, students' task performance and the accuracy of their self-assessment may be influenced by the fact that they have to select relevant criteria. This study explores how selecting criteria affects perceived cognitive load as well as task performance and accuracy of self-assessment.

The main goal of this study was to investigate how drawing students' attention to relevant assessment criteria influences their task performance and self-assessment skills. For one group of students the relevant criteria for specific tasks were highlighted in the whole set of criteria, while another group of students was given an undifferentiated list of all relevant criteria. The first hypothesis was that students who are shown which criteria are relevant show higher task performance than students who are given all criteria, because the former have a better idea of what good task performance should look like (Dochy et al. 1999). The second hypothesis was that students who have practised self-assessment skills with relevant criteria do better at selecting relevant criteria during a test phase than students who practised with an undifferentiated full set of criteria. The third, related, hypothesis was that students who have practised with relevant criteria are less likely to select irrelevant criteria during a test phase. We also explored the difference in perceived cognitive load between students who self-assessed performance with relevant criteria and students who self-assessed with the full list of criteria. Because selecting relevant performance criteria is the first step in a cyclical process of learning how to self-assess (see Figure 1), the fourth hypothesis was that students who have practised with relevant criteria judge their performance more accurately than students who have practised with all criteria. Finally, to close the cyclical process of self-assessment, the fifth hypothesis was that students who have practised with relevant criteria are better able to formulate points of improvement than the students who have practised with the full list of criteria, because the support they have received in identifying relevant strengths and weaknesses helps them to set appropriate learning goals for future tasks (Eva and Regehr 2005). We also explored differences between nursing and care students with regard to task performance and self-assessment skills. The nursing students participating in this study were at level 4 of the European Qualifications Framework, i.e. they had to acquire factual and theoretical knowledge in broad contexts within the field of nursing in a four-year study programme. The care students in this study were at level 3, i.e. they had to acquire knowledge of facts, principles, processes and general concepts relating to nursing. Their study programme lasts three years. The major difference between the programmes is the stronger theoretical orientation of the nursing programme. The programmes have

almost identical first years. As professional boundaries fade and healthcare professionals of various levels have to perform the same tasks, they will be judged against the same set of criteria (Fotheringham 2010). The question is therefore relevant whether these two groups of students need different types of support, because of the different orientations of their study programmes.

Method

Participants and design

First-year students of secondary vocational education programmes in nursing and care ($N = 68$; 6 males and 62 females) participated in the study. The programmes consist of formal education and working in real practice. The study used a 2 x 2 factorial design, with Relevance (Relevant Criteria vs All Criteria) and Programme (Nursing vs Care) as between-subjects factors. Students from the two programmes were randomly assigned to one of two Relevance conditions, i.e. a condition in which they were given the relevant criteria for a task or a condition in which they received the full set of criteria without being informed which criteria were relevant to a particular task. There were 18 students in the relevant criteria/nursing group, 18 students in the all criteria/nursing group, 16 students in the relevant criteria/care group, and six students in the all criteria/care group. The data of 10 students in the all criteria/care group were lost due to a technical problem with the electronic learning environment, which means that the results for this group should be interpreted with care.

Learning materials

The electronic learning environment 'Care Village' was used to collect data for the study. Care Village presents an authentic work environment in which all relevant care and nursing settings are available (e.g. hospital, psychiatric hospital, care for the elderly; Gulikers et al. 2008). Students have access to a virtual school and a multimedia centre, and can select learning tasks for a care setting of their choice. No distinction was made for nursing and care students. For the study, three different types of learning tasks were developed by a project group consisting of experts and teachers in the field of care and nursing, according to the principles of the four-component instructional design model (van Merriënboer 1997; van Merriënboer and Kirschner 2007): (1) worked-out examples in which students were instructed to study a task performed by someone else and answer questions about it; (2) completion tasks requiring students to perform part or parts of a task; and (3) conventional tasks requiring students to perform the whole task independently. Each learning task consisted of a case description, leading questions to help students understand the case, an assignment to be performed in school and an assignment to be performed in a practice setting. Students first read the task description in the electronic learning environment Care Village and then performed the assignments. For most learning tasks, the assignment involved providing nursing care to a simulated patient in a simulated setting (school) and in real life (workplace), and completing a nursing record.

After finishing a particular learning task, students self-assessed their performance and generated points of improvement for the next learning task. A comprehensive set of 69 assessment criteria was developed by a project group consisting of expert

teachers with a background in nursing. This set of criteria covered the full set of tasks a nurse has to perform at novice level. The assessment criteria were made operational in scoring rubrics which students used to indicate their competence level for each relevant criterion (Sluijsmans, Straetmans, and van Merriënboer 2008).

Students in the two experimental groups could choose from the same set of learning tasks and were given the same set of 69 assessment criteria. The relevant criteria groups received a set in which the relevant criteria for the learning task were in **bold** print and the irrelevant criteria in normal print; the all criteria groups received a list in which all criteria were in **bold** print. The relevant criteria groups thus received all criteria but were alerted to the relevance or irrelevance of the criteria, which enabled them to use relevant criteria in self-assessing their performance. The students in the all criteria groups had to determine for themselves which criteria were relevant and use these to self-assess their performance. The latter condition was identical to that of the regular study programme. For the study, all students had to self-assess their performance on all learning tasks during a four-month period. A teacher also assessed the students' task performance using the list of relevant criteria. The virtual school of Care Village allowed students to compare their self-assessments with the teacher's assessments.

After the four-month period, the students had to perform a test task. This was a conventional task resembling the learning tasks. All students were given the list of 69 assessment criteria with no indication of relevance or irrelevance, and they were asked to select the criteria they thought relevant and use these to assess their performance on the test task, which consisted of the school-based part of the task. All students' actions in Care Village (e.g. number and nature of learning tasks performed, self assessments, teacher assessments) were logged automatically.

Measurements

Selection of relevant criteria

Correct selection of relevant criteria from the list of 69 criteria was considered indicative of students' competence in selecting relevant criteria for a particular learning task. It was measured by comparing the criteria selected by the student with the relevant criteria identified by the expert teachers. The number of true positives selected by a student was calculated and divided by the total number of potentially relevant criteria (i.e. true positives), yielding a minimum score of 0 and a maximum score of 1, the latter showing that the student had selected all true positives.

Selection of irrelevant criteria

The number of irrelevant criteria selected for self-assessment was considered to indicate students' competence in identifying which criteria are irrelevant for a particular learning task, i.e. in not choosing true negatives. This was measured by comparing the criteria not selected by the student with the irrelevant criteria identified by the expert teachers. The number of true negatives selected by a student was divided by the total number of irrelevant criteria (i.e. true negatives). A minimum score of 0 and a maximum score of 1 could be obtained, the latter indicating that the student had selected no false positives.

Accuracy of self-assessment

The accuracy of self-assessment was determined by the agreement between student self-assessment and teacher assessment of the same task. According to Boud and Falchikov (1989), expert teachers know better than novice students when performance meets predefined criteria and thus how to assess student performance. Fowles (2009) has described several ways of measuring agreement between different types of judges, such as teacher–teacher agreement and student–teacher agreement. A method for measuring student–teacher agreement described by Rust, Price, and O'Donovan (2003) uses a simple numerical system to compare scores of teachers and students, with 0 indicating the same score, 1 indicating a one-grade difference (+1 if the student's grade is higher, −1 if the student's grade is lower), 2 indicating a two-grade difference, etc. As we were not interested in the direction of the difference between student and teacher judgements, we used the absolute value of the difference as a measure of the accuracy of self-assessment, with an optimal score of 0. This approach is justified by Murphy's suggestion (1982) that positive and negative grade differences can result in a misleading low mean grade difference.

Generating points of improvement

For each learning task, students could suggest points for improvement based on their self-assessment on a particular criterion. A score of 1 was given when a student suggested one or more points for improvement; otherwise, a score of 0 was assigned.

Learning task performance

Learning task performance indicates how well a student executed a learning task. The teachers received the list with only the relevant criteria for assessing task performance, as was customary for regular assessments. The overall score on learning task performance was the average score on all relevant criteria. Criteria were rated on a four-point scale. In determining students' final grades after the experiment, a correction was made for each experimental group to ensure that participation in the experiment had no negative effects for the participating students. This was done by calculating the mean difference between the means of the students in the relevant criteria groups and in the all criteria groups. The mean difference was added to the individual grades of the students in the all criteria groups. Furthermore, students in the all criteria groups were offered additional training in areas where teachers had identified a need for more learning experiences.

Test task performance

Performance on the test task was measured in the same way as learning task performance, using the same 4-point rating scale. Two independent expert nurses who were not members of the teaching staff assessed task performance. Interrater-reliability was high (Pearson $r = .92$, $p = .01$); the mean absolute difference between the two raters was .31 ($SD = .31$).

Self-assessment of mental effort. After self-assessing their learning tasks and the test task, students were asked to rate the 'mental effort required to perform the

self-assessment' on a 7-point cognitive-load rating scale similar to the one used in an experiment in secondary vocational education by Corbalan, Kester, and van Merriënboer (2009) (1 = very little mental effort; 7 = very high mental effort).

Student perceptions. Student perceptions of the following 7 aspects were measured using a four-point Likert scale: Relevance of self-assessment, ability to self-assess, interesting course material, task orientation, interest in and enjoyment of the learning tasks and reflection and usefulness. The self-directed learning skills questionnaire of Kicken, Brand-Gruwel and van Merriënboer (2006) was used to measure student perceptions of the relevance of self-assessment and their ability to self-assess. Two scales (interesting course material and task orientation) of the Inventory of Perceived Study Environment (IPSE; Wierstra et al. 1999) were used to measure student perceptions regarding the learning environment, consisting of Care Village and the non-electronic aspects of the actual learning environment. Three scales of the Intrinsic Motivation Inventory developed by Deci et al. (1994) and translated into Dutch by Martens and Kirschner (2004), were used to measure interest in and enjoyment of the learning tasks and reflection and the usefulness of the learning environment. In this study, Cronbach's alpha scores of the perception scales ranged from .66 to .93, which can be considered to be acceptable or high.

Procedure

At the start of the experiment, students and teachers received written and verbal instructions on working in Care Village. During a period of four months (learning phase), students worked on the learning tasks, self-assessed their performance on each task and rated the mental effort they had to expend during self-assessment. At the end of the experiment, students performed the test task, self-assessed their performance on that task and rated their mental effort in the same way (test phase). Finally, students were asked to complete the 7 student perception scales.

Results

This section describes the results for the dependent variables obtained during the learning phase and the test phase and students' perceptions. Two-way ANOVAs were conducted to test for effects of Relevance (relevant vs all criteria) and Programme (nursing vs care). For all analyses, the significance level was set at .05. Partial eta-squared is provided as a measure of effect size, with $\eta_p^2 = .01$ corresponding to a small effect, $\eta_p^2 = .06$ to a medium effect, and $\eta_p^2 = .14$ to a large effect (Kittler, Menard, and Phillips 2007). All analyses were performed using SPSS 15.0 for Windows.

Table 1 presents the means and standard deviations for the dependent variables in the test and learning phase.

Test phase

The first hypothesis regarding test performance was confirmed by a main effect of Relevance ($F(1, 54) = 3.750$, $MSE = 1.890$, $p = .028$, $\eta_p^2 = .065$), indicating that students in the relevant criteria groups ($M = 1.88$, $SD = .83$) outperformed students in the all criteria groups ($M = 1.70$, $SD = .62$). Also a main effect of Programme was found ($F(1, 54) = 7.435$, $MSE = 3.747$, $p = .009$, $\eta_p^2 = .121$), indicating that

Table 1. Means and standard deviations of measures during the test and learning phase.

	All Criteria Group (n = 24)				**Relevant Criteria Group (n = 34)**			
	Nursing (n = 18)		Care (n = 6)		Nursing (n = 18)		Care (n = 16)	
	M	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
Test Phase								
Test task performance	1.93	.49	1.03	.48	2.00	.69	1.76	.96
Selection of relevant criteria	.30	.32	.50	.34	.37	.30	.61	.35
Selection of irrelevant criteria	.20	.29	.37	.22	.21	.21	.48	.39
Accuracy of self-assessment	1.07	.90	2.10	.68	.90	1.04	1.23	1.49
Generation of points of improvement	.44	.51	.00	.00	.17	.38	.25	.45
Self-assessment mental effort	2.94	1.31	2.67	1.21	3.51	1.46	3.50	1.21
Learning Phase								
Learning task performance	3.61	.32	2.92	.32	3.58	.40	2.61	.60
# Tasks finished during learning phase	2.94	3.51	3.00	4.20	4.89	3.66	2.25	3.32
Accuracy of self-assessment	.56	.42	.72	.49	.84	.57	.63	.39
Self-assessment mental effort	3.11	.84	3.14	.63	3.30	1.04	2.85	1.18

students in the nursing groups (M = 1.96, SD = .59) outperformed students in the care groups (M = 1.56, SD = .91). No interaction effect was found.

A main effect of Programme was found for selecting relevant criteria (F(1, 54) = 5.407, MSE = .561, p = .024, η_p^2 = .091), indicating a higher proportion of selected relevant criteria for the care groups (M = .58, SD = .34) than for the nursing groups (M = .34, SD = .31). The results for selecting relevant criteria did not confirm the second hypothesis as no effect was found for Relevance and no interaction effect was found.

A main effect of Programme was found for selecting irrelevant criteria (F(1, 54) = 7.068, MSE = .604, p = .01, η_p^2 = .116), indicating a higher proportion of selected irrelevant criteria for the care groups (M = .45, SD = .35) than for the nursing groups (M = .20, SD = .25). The third hypothesis was not confirmed for selecting irrelevant criteria as neither an effect of Relevance nor an interaction effect was found.

The fourth hypothesis was not confirmed as no significant effects were found for accuracy of self-assessment; the mean absolute difference between the students and the expert nurses was 1.44 (SD = .79).

No main effects of Relevance and Programme were found for generating points of improvement, but a significant interaction effect (F(1, 54) = 4.445, MSE = .819, p = .040, η_p^2 = .076) partly confirmed the fifth hypothesis. Visual inspection of this interaction effect in Figure 2 shows that only the care group benefited from being provided with the relevant criteria during the learning phase.

A marginally significant main effect of Relevance was found for perceived mental effort used in self-assessment (F(1, 54) = 3.301, MSE = 5.783 , p = .075, η_p^2 = .058). The relevant criteria groups (M = 3.51, SD = 1.33) indicated higher mental effort for self-assessments than the all criteria groups (M = 2.88, SD = 1.26).

Learning phase

On average, students completed 3.36 learning tasks during the learning phase with a standard deviation of 3.64. Variation was high, with a minimum of zero and a maximum of 14 learning tasks.

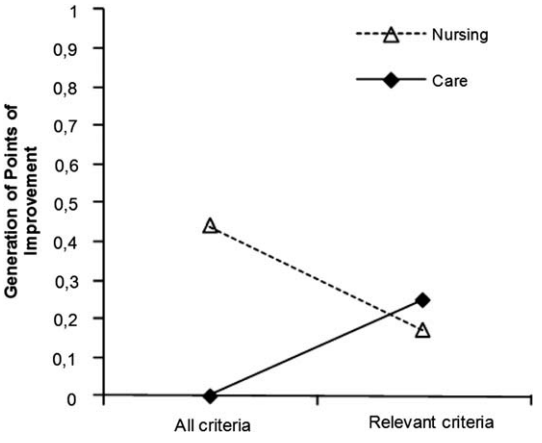

Figure 2. Interactions between the factors Relevance and Program in the test phase, on generation of points of improvement.

A main effect of Programme was found for learning task performance ($F(1, 54) = 42.416$, $MSE = 8.120$, $p = .00$, $\eta_p^2 = .44$). The care group ($M = 2.69$, $SD = .55$) showed lower task performance than the nursing group ($M = 3.59$, $SD = .36$). This suggests that nurses and care students differ in their needs for instructional support. Relevance and learning task performance showed no interaction effects.

No significant effects were found for Relevance, Programme or their interaction in relation to number of tasks completed, accuracy of self-assessment and reported mental effort for self-assessment.

Student perceptions

Table 2 presents the means and standard deviations for the perception measures.

A main effect of Relevance was found for relevance of self-assessment ($F(1, 54) = 4.567$, $MSE = 1.874$, $p = .037$, $\eta_p^2 = .078$), indicating that students in

Table 2. Means and standard deviations for perception measures.

	All Criteria Group (n = 24)				**Relevant criteria Group (n = 34)**			
	Nursing ($n = 18$)		Care ($n = 6$)		Nursing ($n = 18$)		Care ($n = 16$)	
	M	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
Self-Directed Learning Skills								
Relevance of self-assessment	3.00	.66	3.17	.41	2.56	.69	2.81	.63
Ability to self-assess	2.77	.41	2.83	.13	2.74	.43	3.17	.47
Perceived Study Environment								
Interesting course material	2.42	.47	2.79	.33	2.73	.32	2.83	.57
Task orientation	2.35	.71	2.78	.34	2.65	.46	2.98	.56
Intrinsic Motivation								
Interest and pleasure in learning tasks	2.44	.53	2.83	.28	2.74	.38	2.77	.69
Interest and pleasure in reflection	1.39	.45	2.44	.65	1.87	.64	2.49	.49
Usefulness	2.44	.67	2.96	.53	2.88	.50	3.06	.59

the all criteria groups (M = 3.04, SD = .60) perceived the self-assessments as more relevant than students in the relevant criteria groups (M = 2.68, SD = .67).

A main effect of Programme was found for self-assessment ability ($F(1, 54)$ = 4.140, MSE = .723, p = .047, η_p^2 = .071), indicating that perceived self-assessment ability was higher in the care groups (M = 3.08, SD = .43) than in the nursing groups (M = 2.76, SD = .41).

As for interesting course material, no significant differences between conditions were found.

A main effect of Programme was found for task orientation ($F(1, 54)$ = 5.174, MSE = 1.684, p = .027, η_p^2 = .087), indicating that students in the care groups (M = 2.92, SD = .51) perceived they had a clearer view of what was expected from them in the learning tasks compared to students in the nursing groups (M = 2.50, SD = .61).

A main effect of Programme was found for enjoyment of and interest in reflection ($F(1, 54)$ = 27.753, MSE = 8.243, p = .00, η_p^2 = .339), indicating that students in the care groups (M = 2.48, SD = .52) perceived they derived more enjoyment from and were more interested in reflection than students in the nursing groups (M = 1.63, SD = .60).

Finally, a main effect of Programme was found for usefulness ($F(1, 54)$ = 4.228, MSE = 1.446, p = .045, η_p^2 = .073), indicating that students in the care groups (M = 3.03, SD = .56) perceived the tasks as more useful than students in the nursing groups (M = 2.66, SD = .62).

Discussion

The goal of this study was to investigate whether and how directing students' attention to relevant performance criteria had an effect on task performance and on the development of self-assessment skills. The first hypothesis, stating that students who are alerted to relevant criteria show higher test task performance than students who receive an undifferentiated list of criteria, was confirmed by our findings. The finding that students who were given relevant criteria were better task performers validates the notion of Dochy et al. (1999) that providing students with relevant criteria improves their understanding of good task performance, which in turn helps them to perform better. The finding that nursing students outperformed care students on task performance confirms the results of the learning phase, where nursing students also outperformed care students. This may be explained by the different levels of the study programmes of the two groups. Nursing students presumably have a higher level of cognitive ability and consequently do better on task performance.

The second hypothesis, stating that students who are given relevant criteria for self-assessment are better able to select relevant criteria during the test phase than students who practise without being given relevant criteria was not confirmed by the data. A possible explanation is that the acquisition of a complex skill like self-assessment takes more time than was available in this study (van Merriënboer and Kirschner 2007). The results showed that care students selected more relevant criteria than nursing students.

The third hypothesis, stating that students who receive relevant criteria during the learning phase are less likely to select irrelevant criteria during the test phase than students who are given the whole set of criteria, was not confirmed by the data. A possible explanation is that highlighting relevant criteria focuses students'

attention on the relevant criteria only. Students may have assessed their performance using the relevant criteria and simply ignored the irrelevant criteria during practice. It may be possible to obtain a stronger effect by emphasising the reasons underlying the relevance of criteria during practice, for example, by asking students why some criteria are relevant and others are not. The results show that care students selected more irrelevant criteria than nursing students. Combined with the results on the selection of relevant criteria, it can be concluded that care students choose more criteria overall without paying attention to their relevance.

The fourth hypothesis, stating that students who are given the relevant criteria show more accurate self-assessments than students who receive the whole set of criteria, was not confirmed neither during the learning phase nor during the test phase. This is in contradiction with our expectations but in line with findings of Dunning et al. (2004), who also found that knowledge of relevant criteria does not necessarily imply that novice students are able to self-assess their performance on those criteria. More intensive practice may be required to achieve this.

The fifth hypothesis, stating that students who receive relevant criteria generate more points of improvement than students who receive all criteria, was partly confirmed by our findings. Care students profited more from being provided with relevant criteria than nursing students, with highlighting of relevant criteria resulting in more points of improvement in the care group but not in the nursing group. However, more points of improvement did not result in better test task performance of the care students. Apparently, they were not able to use their points of improvement to actually improve their performance.

As for cognitive load, our results show no difference between the groups during the learning phase, but students who received relevant criteria during this phase reported higher mental effort for self-assessment during the test phase than students who received all criteria. Students who practised with relevant criteria may have been more engaged in identifying relevant criteria during the test phase and therefore have invested more effort in self- assessment, with a positive load for learning (Sweller et al. 1998). Another possible explanation is that students who had practised with relevant criteria experienced the assessment task in the test phase as more difficult compared to the students in the all criteria group. For the former group, the test phase was the first time they were confronted with the undifferentiated list of criteria. In this case, the load would be negative for their learning (Sweller et al. 1998).

An intriguing finding is that students in the all criteria groups perceived the self-assessments as somewhat more relevant compared to students in the relevant criteria groups, possibly because they had to actively consider which criteria were relevant and which were not. Furthermore, the care students were more positive in their perceptions, with higher perceived ability with regard to self-assessment. They indicated that they had a clearer task orientation, experienced more enjoyment of and were more interested in reflection, and they perceived the learning tasks as more useful. Overall, they seemed more positive than the nursing students, although this was not reflected in their task performance.

There are two practical implications of the results of our study. Firstly, in order to improve task performance, teachers should provide novice students not only with a list of possibly relevant performance criteria but also clearly point out which criteria are relevant for which task as this helps students to better perform the tasks. Secondly, it seems fruitful to make a distinction between students with relatively low and relatively high performance on learning tasks (in this study the care

students and the nursing students, respectively). Pointing out relevant criteria seems especially important for low-performing students, because it has positive effects on generating points of improvement to be addressed in future tasks. Thus, especially for low-performing students, it is important to know in advance which criteria are relevant and to be trained in how to select relevant criteria.

Future research should investigate how the provision of relevant criteria can be combined with other instructional measures in order to improve the accuracy of self-assessments. Students could be given exemplars of good performance (e.g. video models) to make the relevant criteria for assessment more concrete and explicit. This approach might also be effective in improving students' skills in avoiding the selection of irrelevant criteria. When students are not only confronted with abstract criteria but also with good exemplars, it should become clearer to them what is relevant and what is irrelevant for assessing a particular task. In addition to drawing students' attention to relevant criteria, it would be interesting to explain to them why particular criteria are relevant and others are not, or to ask students to explain this (i.e. 'self-explanation'; Renkl 2002). Thinking-aloud protocols in which students verbalise their thoughts during self-assessment may also provide insight into the difficulties encountered by students and the type of cognitive load they experience.

To conclude, the results of this study indicate that it is worthwhile to point out to students which assessment criteria are relevant for particular learning tasks. Highlighting these criteria enables students to perform better and to practise self-assessment skills with the appropriate criteria. Care students especially appear to profit from highlighting relevant criteria, with positive effects on generating points of improvement. Further research is needed to develop additional guidelines for improving the accuracy of self-assessments and making students aware of the difference between relevant and irrelevant criteria.

References

Boud, D. 1999. Avoiding the traps: Seeking good practice in the use of self assessment and reflection in professional courses. *Journal of Social Work Education* 18: 121–32.

Boud, D., and A. Brew. 1995. Developing a typology for learner self assessment practices. *Research and Development in Higher Education* 18: 130–35.

Boud, D., and N. Falchikov. 1989. Quantitative studies of student self-assessment in higher education: A critical analysis of findings. *Higher Education* 18: 529–49.

Boud, D., and N. Falchikov. 2006. Aligning assessment with long-term learning. *Assessment and Evaluation in Higher Education* 3: 399–413.

Corbalan, G., L. Kester, and J.J.G. van Merriënboer. 2006. Towards a personalized task selection model with shared instructional control. *Instructional Science* 34: 399–422.

Corbalan, G., L. Kester, and J.J.G. van Merriënboer. 2009. Selecting learning tasks: Effects of adaptation and shared control on efficiency and task involvement. *Contemporary Educational Psychology* 33, no. 4: 733–56.

Deci, E.L., H. Eghrari, B.C. Patrick, and D. Leone. 1994. Facilitating internalization: The self-determination theory perspective. *Journal of Personality* 62: 119–42.

Dochy, F., M. Segers, and D.A. Sluijsmans. 1999. The use of self-, peer and co-assessment in higher education: A review. *Studies in Higher Education* 24: 331–50.

Dunning, D., C. Heath, and J.M. Suls. 2004. Flawed self-assessment: Implications for health, education, and the workplace. *Psychological Science in the Public Interest* 5, no. 3: 69–106.

Eva, K.W., and G. Regehr. 2005. Self-assessment in the health professions: A reformulation and research agenda. *Academic Medicine* 80, no. 10: 46–54.

Falchikov, N., and D. Boud. 1989. Student self-assessment in higher education: A meta-analysis. *Review of Educational Research* 59: 395–430.
Fotheringham, D. 2010. Triangulation for the assessment of clinical nursing skills: A review of theory, use and methodology. *International Journal of Nursing Studies* 47: 386–91.
Fowles, D. 2009. How reliable is marking in GCSE English? *English in Education* 43, no. 1: 50–67.
Gulikers, J.T.M., T.J. Bastiaens, P.A. Kirschner, and L. Kester. 2008. Authenticity is in the eye of the beholder: Student and teacher perceptions of assessment authenticity. *Journal of Vocational Education & Training* 60: 401–12.
Jourdain, G., and D. Chênevert. 2010. Job demands-resources, burnout and intention to leave the nursing profession: A questionnaire survey. *International Journal of Nursing Studies* 47: 709–22.
Kicken, W., S. Brand-Gruwel, and J.J.G. van Merriënboer. 2006. *Self-directed learning skills questionnaire*. Heerlen: Open Universiteit Nederland.
Kicken, W., S. Brand-Gruwel, and J.J.G. van Merriënboer. 2008. Scaffolding advice on task selection: A safe path toward self-directed learning in on-demand education. *Journal of Vocational Education and Training* 60: 223–39.
Kicken, W., S. Brand-Gruwel, J.J.G. van Merriënboer, and W. Slot. 2009. The effects of portfolio-based advice on the development of self-directed learning skills in secondary vocational education. *Educational Technology Research and Development* 57, no. 4: 439–60.
Kittler, J.E., W. Menard, and K.A. Phillips. 2007. Weight concerns in individuals with body dysmorphic disorder. *Eating Behaviors* 8: 115–20.
Kruger, J., and D. Dunning. 1999. Unskilled and unaware of it: How difficulties in recognizing one's own incompetence lead to inflated self-assessments. *Journal of Personality and Social Psychology* 77: 1121–34.
Martens, R.L., and P.A. Kirschner. 2004. What predicts intrinsic motivation? Paper presented at the 2004 AECT convention, October 19–23, Chicago.
Murphy, R.J.L. 1982. A further report of investigations into the reliability of marking of GCE examinations. *British Journal of Educational Psychology* 52: 58–63.
Orsmond, P., S. Merry, and K. Reiling. 2002. The use of exemplars and formative feedback when using student derived marking criteria in peer and self assessment. *Assessment & Evaluation in Higher Education* 27: 309–23.
Regehr, G., and K. Eva. 2006. Self-assessment, self-direction, and the self-regulating professional. *Clinical Orthopaedics and Related Research* 449: 34–8.
Renkl, A. 2002. Worked-out examples: Instructional explanations support learning by self-explanations. *Learning and Instruction* 12: 529–56.
Rust, C., M. Price, and B. O'Donovan. 2003. Improving students' learning by developing their understanding of assessment criteria and processes. *Assessment & Evaluation in Higher Education* 28, no. 2: 147–64.
Sadler, D.R. 1989. Formative assessment and the design of instructional systems. *Instructional Science* 18: 119–44.
Sluijsmans, D.M.A., G.J.J.M. Straetmans, and J.J.G. van Merriënboer. 2008. Integrating authentic assessment with competence-based learning in vocational education: The Protocol Portfolio Scoring. *Journal of Vocational Education & Training* 60: 159–72.
Sweller, J., J.J.G. van Merriënboer, and F. Paas. 1998. Cognitive architecture and instructional design. *Educational Psychology Review* 10: 251–94.
Van Merriënboer, J.J.G. 1997. Training complex cognitive skills. Englewood Cliffs, NJ: Educational Technology Publications.
Van Merriënboer, J.J.G., and P.A. Kirschner. 2007. Ten steps to complex learning. Mahwah, NJ: Erlbaum/Taylor and Francis.
Van Merriënboer, J.J.G., and J. Sweller. 2005. Cognitive load theory and complex learning: Recent developments and future directions. *Educational Psychology Review* 17: 147–77.
Wierstra, R.F.A., G. Kanselaar, J.L. van der Linden, and H.G.L.C. Lodewijks. 1999. Learning environment perceptions of European university students. *Learning Environments Research* 2: 79–98.
Wolf, L.D., P. Potter, J.A. Sledge, S.B. Boxerman, D. Grayson, and B. Evanoff. 2006. Describing nurses' work: Combining quantitative and qualitative analysis. *Human Factors* 48: 5–14.