Improving sustainable assessment skills in vocational education

Improving sustainable assessment skills in vocational education

The research reported here was carried out at the

Open Universiteit
www.ou.nl

in the context of the research school

ICO
Interuniversity Centre for Educational Research

and was funded by

ROC A 12
Regionaal OpleidingsCentrum A12

ISBN 978 90 79447 48 0

Cover design: Wim De Buck
Printing: Datawyse / Universitaire Pers Maastricht

Improving sustainable assessment skills in vocational education

PROEFSCHRIFT

ter verkrijging van de graad van doctor
aan de Open Universiteit
op gezag van de rector magnificus
prof. mr. A. Oskamp
ten overstaan van een door het
College voor promoties ingestelde commissie
in het openbaar te verdedigen

op vrijdag 11 maart 2011 te Heerlen
om 13:30 uur precies

door
Greet Mia Jos Fastré
geboren op 16 april 1981 te Hasselt

Promotor

Prof. dr. J.J.G. van Merriënboer, Open Universiteit

Co-promotor

Dr. M.R. van der Klink, Open Universiteit

Overige leden beoordelingscommissie

Prof. dr. D. Gijbels, Universiteit van Antwerpen
Prof. dr. A.F.M. Nieuwenhuis, Universiteit Twente
Prof. dr. L.W.T. Schuwirth, Universiteit Maastricht
Prof. dr. M.S.R. Segers, Universiteit Maastricht
Prof. dr. R.L. Martens, Open Universiteit

VOORWOORD

In september 2006 begon ik aan een reis waar ik al enkele jaren naar uitkeek. Eindelijk kreeg ik die promotieplek die ik zo begeerde. Het werd een boeiende en leerzame reis waarin heel wat bergen beklommen werden. Ik wil dan ook graag een aantal mensen bedanken zonder wie ik deze reis nooit tot een goed einde had kunnen brengen.

In de eerste plaats wil ik mijn promotor Jeroen van Merriënboer en mijn dagelijks begeleider Marcel van der Klink bedanken. Jeroen en Marcel, tijdens onze laatste lange autorit naar Ede zei ik jullie al dat ik ons een fantastisch team vond. En dat meende ik. In de afgelopen jaren heb ik enorm genoten van onze samenwerking en die ga ik ook heel erg missen! Jeroen, gedurende de vier jaar van mijn proefschrift, was jij één van de weinige constante factoren. Je bent er altijd voor je aio's en dat siert je. Als ik je vroeg om een planning te maken, dan gaf je als antwoord: 'Planning wat is dat: doorwerken!' Maar daarnaast hield je me overeind door altijd te beloven: 'Het komt wel goed.' En hier sta ik dan, vier jaar later. En het is goed gekomen! Marcel, jij kwam iets later in mijn traject maar vormde de perfecte aanvulling voor ons team! Met spreuken als: 'Tom Poes verzin een list' en 'Met vlag en wimpel, houd het simpel' wist jij mijn onderzoek mee over de eindstreep te trekken. Ik vond het heel spannend toen jij mijn begeleider werd omdat ik je niet goed kende, maar je was een erg betrokken en fantastisch leuke begeleider die me de juiste ondersteuining wist te geven en altijd voor me klaar stond. Bedankt hiervoor!

Dit onderzoek zou er niet geweest zijn zonder de steun van ROC A12. Bedankt om de noodzaak van dit project in te zien. Buiten de financiering, stond er steeds een geweldig docententeam klaar. Zonder hen had ik dit onderzoek niet kunnen bedenken. In het bijzonder wil ik graag Ellen Leenaerts, Lidwien de Nijs en Pauline Amsing-Smit bedanken die elk op hun eigen manier fantastische projectleiders waren. Een bijzonder woord van dank aan Pauline, met wie ik de belangrijkste onderzoeken heb opgezet en bedacht. Wij hadden maar een paar woorden nodig om elkaar te begrijpen en dat was erg prettig en efficiënt! Ook wil ik graag het ontwikkelteam en alle studenten die meewerkten aan mijn onderzoek bedanken. Voor het laatste onderzoek mochten we ook een kijkje nemen in de keuken van ROC RIvor en het Graafschapcollege. Ook voor jullie een grote dankjewel voor de medewerking aan mijn onderzoek.

Voor de technische ondersteuning en het opzetten van mijn mooie leeromgevingen wil ik graag Bas Janssen, Martin Schoenmaker en Marco van Hout van Moni-

to bedanken. Voor het geweldig leuke filmwerk wil ik graag Maurice Copier en Frans Bronzwaer van In Beeld en de filmploegen bedanken. Het was leuk om met jullie samen te werken! Dankjewel Wim De Buck voor het ontwerp van mijn kaft. Verder wil ik graag Nicole Knebel, Ingrid Jonkman en Audrey Wigman bedanken voor de nodige ondersteuning bij mijn onderzoeken. Dankjewel Mieke Haemers voor het kritisch nalezen van mijn proefschrift. Zoveel details op zo een korte tijd, echt knap!

Daarnaast wil ik graag alle aio's en ex-aio's van de OU bedanken voor de gezellige tijd die we samen hadden. In het bijzonder Gemma, je was mijn buurvrouw en bent nu zo ver weg. Maar je bent nog steeds mijn lieve amiga en ik zal je nooit vergeten. Quando piensa a Gemma… Bettine, je kwam pas later in mijn traject op mijn pad maar er was meteen een klik die zorgde voor heel wat leuke momenten. Helen, onze gedeelde liefde voor mooie juweeltjes en jouw vrolijkheid zullen me nog lang bij blijven. Ook zou ik graag een aantal andere collega's van de OU willen bedanken. Iwan, voor de vele gezellige koffiemomentjes. Ellen, voor je warme persoonlijkheid en je lieve bezorgdheid. En onze lieve secretaresse Ingrid die altijd tijd maakte voor een babbeltje als het werk even te veel werd.

Ook heb ik de eer gehad een deel van mijn traject op de UM te mogen verblijven. Dankjewel voor het warme onthaal en de leuke tijd die ik daar had. In het bijzonder wil ik mijn kamergenootjes Janneke, Floris, Jeantine, Rachelle, Marjo en Ellen bedanken voor de gezamenlijke 'even geen zin om te werken' momenten, de gezellige gesprekken en de leuke lunches. Jullie kwamen in mijn leven in een periode dat er grote dingen gebeurden en het was fantastisch om die met jullie te kunnen delen!

Danny en Marjo, mijn paranimfen. Wat vind ik het fijn dat jullie vandaag achter mij staan. Lieve Marjo, ook jij was één van de weinige factoren die constant bleef in mijn aio-tijd. Ik vond je een erg leuk kamergenootje en vond het altijd heerlijk om met jou sappige roddels te kunnen delen, te lachen en lief en leed te doorstaan. We hebben allebei een turbulente 4 jaar achter de rug en het was fijn dat we er altijd voor elkaar konden zijn! Ik ga je missen.

Lieve Danny, toen ik je pas leerde kennen moest ik even aan je wennen maar wat ben ik je gedurende de jaren gaan waarderen als een superlieve, betrokken en leuke collega! We gingen door dezelfde fasen in ons onderzoek en konden over alles met elkaar praten. Onze topmomenten waren de lange lunchwandelingen waarin we onze zorgen en pleziertjes met elkaar deelden. Dankjewel dat je er was en ik ga je missen nu je in het hoge noorden woont.

Mijn vriendinnetjes Pascale, Katrien en Kim voor hun interesse in mijn onderzoek en vooral voor alle plezier en afleiding daarbuiten. Pascale en Katrien, toevallig kwam ik in een setting terecht waar jullie ook in thuis waren en kon ik jullie daardoor betrekken bij mijn onderzoek. Dat was heel speciaal voor mij! Pascale, je bent mijn allerliefste vriendinnetje en ik hoop dat dat altijd zo zal blijven!

Dankjewel aan mijn schoonfamilie om te blijven geloven dat er ooit een dag kwam waarop ik zou ‘afstuderen’. Dankjewel aan mijn broer en zus voor de gezellige afleiding en jullie vele vragen over wat ik nu precies deed en vooral wat er nu ging gebeuren op die grote dag.

Een heel bijzonder dankjewel aan mijn ouders. Papa, als jij er niet was, was ik nooit aan deze reis begonnen. Bedankt voor je eeuwige steun en interesse en zelfs om mijn chauffeur en assistent te spelen in ‘verre’ Nederlandse oorden. Mama, jij bent mijn beste steun als ik het even gehad heb met het werk want jij weet altijd wel voor de allerbeste afleiding te zorgen. Dankjewel dat jullie er altijd zijn als ik jullie nodig heb en voor alle kansen die jullie me gegeven hebben.

En dan tenslotte, de twee mannen in mijn leven, mijn lieve echtgenoot Bas en zoontje Casper. Liefste grote man van me, jij bent er altijd als ik je nodig heb. Jij weet perfect wat ik nodig heb als ik er even door zit. Ik vind het heerlijk om mijn leven met jou te delen. Liefste kleine man, je kwam in mijn leven een jaar geleden en maakte alles relatief met die stralende oogjes en glimlach van jou! Wat zou ik toch beginnen zonder jullie? Bedankt voor jullie onvoorwaardelijke liefde en steun. Ik hou van jullie!

Greet Fastré
Hasselt, 11 maart 2011

CONTENTS

CHAPTER 1

General introduction

In today's vocational education, both students in nursing and students in car mechanics have to be able to analyse information, cooperate with others, and communicate with clients. However, the contexts in which they have to demonstrate these competences differ substantially. For example, student nurses must analyse patient files, while car mechanics students must analyse car defects as part of their future job. Thus, in vocational education, students are expected to develop the same competences although they do not have to show the same task performance.

On the one hand, students must be better prepared for their future job; on the other hand, they must also have the capabilities to keep on developing their competences because the labour market quickly changes (COLO, 2008). Vocational education must prepare students to become competent professionals who cope with future changes in their jobs and have higher job mobility during their careers. Competent professionals are competent task performers and continue learning throughout their professional careers. To be able to deal with the constantly changing circumstances they face, students must become self-directed learners who are able to recognize their own learning needs and put them into action (Boud & Falchikov, 2006; Dearnly & Meddings, 2007; Kicken, Brand-Gruwel, van Merriënboer, & Slot, 2009).

In order to become independent and self-directed lifelong learners, students need to develop sustainable assessment skills which enable them to assess their performance and to keep learning throughout life (Boud, 2000). Sustainable assessment skills help students to meet their own future learning needs because they enable them to assess their own learning process and task performance, and to generate goals for future learning and performance (Butler & Winne, 1995).

A prominent skill in sustainable assessment is to identify relevant criteria for performance and learning. Students can become acquainted with criteria by analysing (a) criteria explicitly provided by others (Tan, 2007), (b) criteria implicitly used by others who assess their performance, and (c) criteria implicitly used by others who demonstrate intended task performance (O'Donnovan, Price, & Rust, 2004). A high degree of transparency of criteria is necessary for students to acquire them (Boud, 2000; Sadler, 1989).

In educational practice, however, assessment criteria are often formulated on a holistic level, not describing the desired performance students must be able to show. Above that, in schools for secondary vocational education students are often expected to select themselves the appropriate criteria for the learning tasks they are working on from a long list with all possibly relevant criteria. The question arises if students in senior vocational education, and especially novice students, are able to use these broadly formulated assessment criteria and to select the criteria that are applicable to particular learning tasks. This dissertation examines what kind of assessment criteria novice students in the domain of Nursing and Care need in order to develop sustainable assessment skills and to become competent professionals.

Overview of the Dissertation

Chapter 2 discusses the importance of sustainable assessment skills, which enable students to assess their own performance and to continue learning throughout life. It presents a model for developing sustainable assessment skills consisting of three components: (a) conditions necessary for developing sustainable assessment skills, (b) constituent parts of sustainable assessment skills, and (c) educational methods for guiding students in their development of sustainable assessment skills. In the three empirical studies following this chapter (for an overview see Table 1.1), the focus is on the development of students' sustainable assessment skills by concentrating on the importance of transparent assessment criteria.

Table 1.1: Schematic overview of the empirical studies.

	Independent variables	**Dependent variables**
Study 1 Chapter 3	Relevance of criteria: - Relevant criteria - All criteria	Task performance Accuracy of self-assessment skills Mental effort
Study 2 Chapter 4	Type of criteria: - Performance-based criteria - Competence-based criteria	Task performance Quality of self-assessment skills Quality of assessment skills Mental effort Instructional efficiency
Study 3 Chapter 5	Relevance of criteria, type of criteria, and their interactions - Competence-based/all criteria - Competence-based/relevant criteria - Performance-based/all criteria - Performance-based/relevant criteria	Task performance Quality of self-assessment skills Quality of assessment skills Mental effort Instructional efficiency

The study reported in Chapter 3 investigates the effects of drawing students' attention to relevant assessment criteria on their performance and self-assessment skills. Two conditions in which students received either the relevant assessment criteria only or all possibly relevant assessment criteria were compared. In the relevant criteria condition, students received the complete list with all possible assessment criteria but the relevant criteria for the task at hand were highlighted. In the all criteria condition, students received the identical list with all possible assessment criteria but without highlights. The students in this condition had to decide independently which criteria were relevant for the task at hand. Students in the relevant criteria condition are expected to perform higher on care and nursing tasks, to make higher quality self-assessments, and to be better able to formulate points of improvement than students in the all criteria condition.

The study presented in Chapter 4 investigates the effects of two different types of criteria on students' performance and self-assessment skills. Two conditions were compared in which students received either performance-based criteria or compe-

tence-based criteria. In the performance-based criteria group, students are provided with a preset list of performance-based assessment criteria describing what students should do for the task at hand. In the competence-based assessment group, students receive a preset list of competence-based assessment criteria describing what students should be able to do. Students in the performance-based condition are expected to perform higher on care and nursing tasks, to experience a lower mental effort during the assessment, and to demonstrate higher quality self-assessments than students in the competence-based group.

Chapter 5 combines the previous studies by investigating the effects of relevance of criteria and type of criteria on students' task performance and self-assessment skills. In this study with a full factorial design, four groups were compared in which students received performance-based/all criteria, performance-based/relevant criteria, competence-based/all criteria, and competence-based/relevant criteria. It is expected that students in the performance-based groups and students in the relevant criteria groups perform higher on care and nursing tasks, experience lower mental effort during the assessment tasks, and make higher quality self-assessments than students in the competence-based and all criteria groups. Furthermore, it is expected that the combination of relevant and performance-based criteria is most beneficial for students.

Finally, Chapter 6 provides an overview of the main findings and conclusions of the studies presented in this dissertation. Theoretical and practical implications are discussed and directions for future research are described.

References

Boud, D. (2000). Sustainable assessment: Rethinking assessment for the learning society. *Studies in Continuing Education, 22*, 151-167.

Boud, D., & Falchikov, N. (2006). Aligning assessment with long-term learning. *Assessment and Evaluation in Higher Education, 31*, 399-413.

Butler, D. L., & Winne, P. H. (1995). Feedback and self-regulated learning: A theoretical synthesis. *Review of Educational Research, 65*, 245-281.

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

Dearnly, C. A., & Meddings, F. S. (2007). Student self-assessment and its impact on learning: A pilot study. *Nurse Education Today, 27,* 333-340.

Kicken, W., Brand-Gruwel, S., van Merriënboer, J. J. G., & Slot, W. (2009). The effects ofportfolio-based advice on the development of self-directed learning skills in secondary vocational education. *Educational Technology Research and Development, 57*(4), 439-460.

O'Donnovan, B., Price, M., & Rust, C. (2004). Know what I mean? Enhancing student understanding of assessment standards and criteria. *Teaching in Higher Education, 9,* 325-335.

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

Tan, K. H. K. (2007). Conceptions of self-assessment: What is needed for long-term learning. In D. Boud & N. Falchikov (Eds.), *Rethinking assessment in higher education: Learning for the longer term* (pp. 181-197). New York: Routledge.

CHAPTER 2

Toward an integrated model for developing sustainable assessment skills

Abstract

One of the goals of current education is to assure that graduates can act as independent lifelong learners. Graduates need to be able to assess their own learning and interpret assessment results. The central question in this chapter is how to acquire sustainable assessment skills, enabling students to assess their performance and learning throughout life, and preparing them to meet their own future learning needs. This chapter presents an integrated model for developing sustainable assessment skills, consisting of three components: (a) conditions necessary for the development of sustainable assessment skills, (b) constituent parts of sustainable assessment skills, and (c) instructional methods for guiding students in the development of sustainable assessment skills. The chapter concludes with suggestions for future research to further develop the proposed model.

This chapter is based on: Fastré, G. M. J., van der Klink, M. R., Sluijsmans, D., and van Merriënboer, J. J. G. (2010). *Toward an integrated model for developing sustainable assessment skills.* Manuscript submitted for publication.

In vocational and professional education, there is increasing recognition that one of the main goals is to educate students to become independent and self-regulated lifelong learners (Boud & Falchikov, 2006). Graduates must be prepared to cope with the challenging demands of rapidly changing professions. The necessity to improve school-to-work transitions and the need to better prepare students for future job demands has stimulated educational institutions to implement competence-based curricula. Such curricula focus on the development of professional skills rather than presenting diplomas based on completed courses, thereby replacing an emphasis on qualifications with an emphasis on capabilities (Biemans, Nieuwenhuis, Poell, Mulder, & Wesselink, 2004). Furthermore, a discipline-based approach is replaced by an integrated approach with a stronger focus on providing authentic learning experiences aimed at the development of competences needed for successful job performance (Tillema, Kessels, & Meijers, 2000).

Competence-based education is expected to lead to a more gradual transition from school to workplace, but it does not necessarily imply that students are prepared for lifelong learning. Being a professional implies not only the ability to perform professional skills, but also the ability to recognize when further improvement of one's own performance or the learning of new skills is required to meet the needs of a rapidly changing profession (McDonald, Boud, Francis, & Gonczi, 1995; Regehr & Eva, 2006). Students should thus learn professional skills as well as assessment skills that help them to recognize their learning needs and, eventually, to improve their performance and learn new skills. For example, student nurses must learn how to wash a patient and communicate with the patient at the same time (professional skills), but they should also recognize that they sometimes stop communicating because they are too much focused on the washing. This might be identified as a point of improvement (assessment skill). According to Boud (2000), students should develop sustainable assessment skills that enable them to assess their performance and to keep on learning throughout life. These assessment skills help students to become self-regulated learners who are aware of their own qualities and shortcomings and who know how to overcome possible hurdles (Butler & Winne, 1995). To accomplish this, they should become responsible for their own learning and be treated as beginning professionals right from the start of their initial professional education (Nieweg, 2004). However, at the start of an educational program, students are often unable to take full responsibility for assessing their own learning and development. The question arises how students can best be guided in the development of sustainable assessment skills.

This chapter starts with explaining the concept of sustainable assessment, followed by a comparison with summative and formative assessment, respectively. Second, an integrated model for the development of sustainable assessment skills is presented. Third, a research agenda is presented to further develop and test the presented model.

What is Sustainable Assessment?

This section gives an overview of the different functions of assessment and compares summative assessment, formative assessment, and sustainable assessment (see Table 2.1). In this overview, the focus is on those aspects of summative and formative assessment that are relevant for explaining the main difference with sustainable assessment. The first, summative, function of assessment focuses on the certification of achievement; grading and pass/fail decisions are the central focus. Summative assessment takes place after a learning phase and serves as assessment of learning (Sadler, 1989). It is a particularly useful function when the transmission of knowledge and skills is the main aim of education. The teacher is typically responsible for student assessment and the student is a passive receiver of grades. Summative assessment is mostly criterion-referenced, meaning that student performance is assessed on the basis of a list with pre-specified criteria (Sadler, 2005; Sluijsmans, Straetmans, & van Merriënboer, 2008). Pure summative assessment gives judgments on these different performance criteria but provides no indication on how to improve performance. The students themselves have to take the responsibility to translate the assessment outcomes into points of improvement for future performance. According to Knight (2002), summative assessment is in disarray because of the negative effects on student learning: It may lead to an overemphasis on grades and learning outcomes and does not properly take the learning process into account. Whereas an important goal of assessment is the certification of students, summative assessment as such is not sufficient and may even be counterproductive to encourage student learning.

Table 2.1: Schematic overview of developments in assessment practices.

	Summative assessment	**Formative assessment**	**Sustainable assessment**
Function	Grading and certification	To aid current learning	To aid lifelong learning
Role division	Teacher: gives grades Student: passively receives grades	Teacher: provides feedback Student: judges performance and receives feedback	Teacher: provides feedback Student: judges performance and actively judges feedback
Criteria	Using preset criteria	Using preset criteria	Developing a critical attitude to criteria; using both preset and newly developed criteria
Scope	Past performance	Current performance	Future performance
Main points of critique	Overemphasis on grades; process blindness	Focus insufficiently on (current) learning; no separation of feedback and grades; insufficient use of self- and peer assessment	Unexplored

The second, formative, function of assessment focuses on the improvement of student performance. A key element is the provision of constant feedback to close the gap between current performance and desired performance, rather than the making of pass/fail decisions. Formative assessment takes place during the learning phase and has the goal of being assessment for learning (Sadler, 1998). In an ideal form of formative assessment, the student has a central role in assessing his or her own performance (i.e., internal feedback) and in processing feedback given by peers and teachers (i.e., external feedback). Feedback is a key element in formative assessment and can be given formally as well as informally during learning (Sadler, 1989). As in summative assessment, in formative assessment, pre-specified criteria are used to assess student performance (Sluijsmans et al., 2008). The results from a summative assessment can be used in a formative way if they are used to inform the students and give feedback on their progress (Dunn & Mulvenon, 2009; Taras, 2005). A comprehensive review study by Black and Wiliam (1998) revealed that formative assessment can lead to significant learning gains, but it also showed substantial weaknesses of current practices. Main problems are that the focus is often not on the learning process but on performance, there is no clear separation of feedback and grading, and there is insufficient use of self-assessment and peer assessment, which are core elements in teaching students to recognize learning gaps. Formative assessment seems promising in theory, but in today's practice, it is insufficient to help students learn to assess their own learning, and it does not prepare them to act as lifelong learners beyond graduation and throughout their entire career.

As a reaction to the dissatisfaction with current formative assessment practices, Boud (2000) introduced sustainable assessment as a third function of assessment. It is "…assessment that meets the needs of the present and prepares students to meet their own future learning needs" (p. 151). Sustainable assessment helps students to become self-regulated learners as sustainable assessment skills help them in the process of planning, monitoring, and controlling their own learning (Boekaerts, Pintrich, & Zeidner, 2000). In order to define sustainable assessment, Boud builds on the principles of formative assessment described in the review by Black and Wiliam (1998). Whereas formative assessment typically focuses on making judgments to improve performance on the next learning task, sustainable assessment focuses not only on performance improvement, but above that on improvement of the learning process in order to stimulate continuous learning beyond graduation. The main function of sustainable assessment is to teach students to self assess their own performance so they can take responsibility for their own professional growth, as this is seen as a critical competence to enable lifelong learning (Tan, 2008). To accomplish this, students judge their own performance not only on the level of single learning tasks but also over a series of tasks, that is, throughout the educational program. The focus of summative and formative assessment is on

current performance and learning whereas sustainable assessment also explicitly focuses on future learning beyond graduation. Furthermore, the concept of sustainable assessment emphasizes that students need to seek actively for external assessments and feedback. Dealing with feedback and reactions to feedback are frequently discussed topics in the formative assessment literature (e.g., Butler & Winne, 1995; Black & Wiliam, 1998; Carless, 2007; Gibbs & Simpson, 2004; Hattie & Timperley, 2007; Shute, 2008), but sustainable assessment goes further in demanding that students make conscious comparisons between self-assessments and assessments by teachers, peers and other stakeholders. The responsibility for the assessment process must gradually shift from the teacher to the students, because, after graduation, students themselves rather than teachers or the curriculum drive the learning process (Boud & Falchikov, 2006; Nieweg, 2004). Like summative and formative assessment, sustainable assessment is typically criterion-referenced. But above that, the concept of sustainable assessment stresses that students also have to develop a critical attitude towards criteria because when they enter the workplace, pre-specified criteria will not always be available to support them in judging their own performance and learning.

Concluding, sustainable assessment builds on the principles of summative and formative assessment to help students becoming assessors of their own learning and consequently self-regulated learners. Although the emerging notion of sustainable assessment seems promising to better prepare students for the rapidly changing workplace, research is needed to explore how it can be applied in educational practice. In the next section, an integrated model for the development of sustainable assessment skills is introduced.

The Model

Figure 2.1 presents the integrated model for developing sustainable assessment skills, including three main components: (a) conditions for developing sustainable assessment skills, (b) constituent parts of sustainable assessment skills, and (c) educational methods for guiding students in the development of sustainable assessment skills. A central component of the model is the description of the constituent parts of sustainable assessment skills (left part of the model) as these are the skills students have to develop to become lifelong learners. The development of these sustainable assessment skills depends on certain conditions that need to be fulfilled, which form the basis for the model (bottom part of the model). The final component describes the educational methods that help students to develop sustainable assessment skills (right part of the model), if the underlying conditions are met. The different components will be discussed in the following sections.

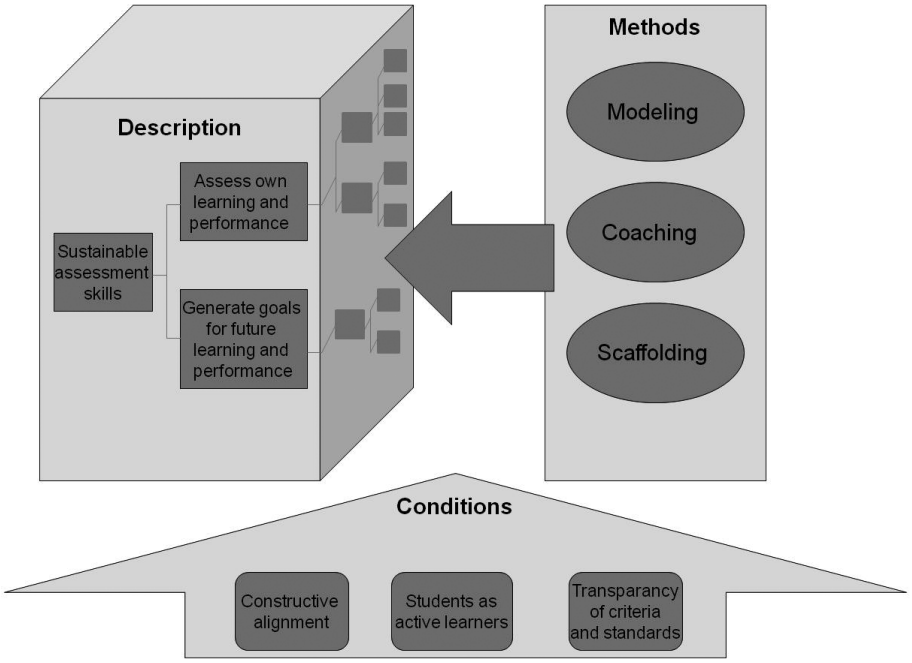

Figure 2.1: The integrated model for developing sustainable assessment skills.

Conditions for Developing Sustainable Assessment Skills

The first component of the model pertains to the conditions necessary for the development of sustainable assessment skills. These conditions are prerequisite to the development of sustainable assessment skills, that is, if the conditions are not met, sustainable assessment skills cannot be developed, irrespective of the educational methods used (Boud, 2000; Boud & Falchikov, 2007). All conditions that are important for formative assessment (e.g., the focus should be on learning rather than performance, reflective assessment with peers should be encouraged) also apply to sustainable assessment. However, for assessment to be sustainable, there are a number of additional conditions which can be classified in three main conditions: (a) there is constructive alignment between instruction and assessments, (b) students are active learners, and (c) performance criteria are transparent.

Constructive alignment

What will be assessed strongly influences what students learn. For example, if students must learn how to communicate with clients, communication skills have to be part of the assessment; if not, students will not be inclined to learn them. It is widely acknowledged that alignment between instruction and assessment is necessary in order to meet the goals of education (Biggs, 1996). Nevertheless, even in vocational and professional education assessment is often treated as an isolated

element of the curriculum. Constructive alignment is most apparent in the self-directed learning cycle, where students go through a cyclical process of performing a task, assessing task performance, identifying points of improvement, and planning future tasks (van Merriënboer & Sluijsmans, 2009). Then, assessment is the link that connects performance on one task to a plan for the following task (Ertmer & Newby, 1996). It ensures that students receive relevant feedback that can be used as 'feed-forward' into future work (Carless, 2007). For example, when a student performs a learning task in which communication with the client is important, and it appears that his communication skills are not meeting pre-defined criteria, the student needs some additional practice and should thus select new tasks that provide precisely this kind of practice.

As indicated before, discontentment with the gap between vocational education and the workplace (Biemans et al., 2004) has encouraged educational institutions to implement learning tasks based on authentic, real-life situations in their vocational curricula. The fruitful use of authentic learning tasks presumes that assessment tasks also resemble real-life professional tasks, because competence-based education and competence-based assessment must go hand in hand (Dochy, Segers, & Sluijsmans, 1999). This does not require assessments always to take place in the workplace or in a genuine social context, but assessment tasks must resemble professional tasks to some degree and the learning they measure should have clear value beyond the classroom (Gulikers, Bastiaens, & Kirschner, 2004; Kerka, 1995). Students must show when they are ready to enter professional practice. For example, a student nurse has to show acceptable performance of nursing tasks with simulated and then real patients, meeting preset performance criteria, before being allowed to enter the labour market.

Boud and Falchikov (2006) advocate that constructive alignment within the program of study is necessary but not sufficient for assessment to be sustainable; alignment with long-term purposes is also of paramount importance. In vocational and professional education, students should develop the skills necessary to recognize their future learning needs, and there should thus be alignment of assessment practices with these future needs (Kicken, Brand-Gruwel, van Merriënboer, & Slot, 2009a). For example, when a new technique is introduced in nursing, graduated nurses should be able to establish their own way of learning to become familiar with this new technique. By integrating an orientation on the future in their programs of study, students get used to this way of thinking, which helps them to become more future-oriented in their careers.

Students as active learners

Feedback on performance is aimed at reducing the discrepancy between current and desired performance. It always addresses three important questions: (1) Where am I going? (2) How am I going? (3) Where to go next? (Hattie & Timperley, 2007).

Obviously, students must learn to interpret external feedback from teachers, peers, or others to determine possible gaps between their actual performance and desired performance (i.e., performance criteria) and to identify points of improvement. Critically considering received feedback requires a safe learning environment (Govaerts, van der Vleuten, Schuwirth, & Muijtjens, 2007). Whatever the feedback is about (e.g., performance, learning, or person-related), it is important that the classroom climate is open so that external feedback can be welcomed and used by students (Liu & Carless, 2006). There has to be ample room for making errors, and feedback should primarily be seen as an instrument for improving performance. In this feedback process, the teacher has an important role in formulating the feedback and creating a safe learning environment.

However, after graduation, it is neither the teacher nor the curriculum but the learners who drive their own learning (Nieweg, 2004). If students only rely on feedback from others, they will never acquire the skills necessary to judge their own performance or to do their own work competently (Boud & Falchikov, 2006). Thus, they must be actively involved in the practice of making informed judgments as part of the curriculum. This may be accomplished through peer assessment and self-assessment, which helps students to gain expertise in the evaluation of task performance (Sadler, 1989). Whereas peer assessment helps students to learn how to formulate feedback on the difference between observed performance and performance criteria, self-assessment eventually enables them to learn how to formulate internal feedback about the difference between their own performance and the criteria. The teacher has to coach the students in acquiring these skills.

Finally, because external feedback can augment, concur, or conflict with internal feedback, students need to develop the ability to cope with the discrepancies between external and internal feedback in order to support their learning (Nicol & Macfarlane-Dick, 2006). For assessment to be sustainable, students need to consciously compare feedback from external parties with their self-assessments (e.g., as part of 360-degree feedback) to obtain an accurate view of their performance and to learn to conduct better self-assessments. For example, a student nurse receives feedback from peers, teachers, senior nurses, and patients during an internship, and the student has to compare that feedback with his own internal feedback and identify discrepancies and possible explanations for these discrepancies. This process will help to clarify performance criteria and to improve the formulation of points-of-improvement. When students would only be passive learners and receive feedback without using it, they could not become lifelong learners. After graduation, students often actively have to seek for feedback and when they receive feedback, for example from clients, they have to be able to deal with it.

Transparency of criteria

When students are not familiar with the performance criteria that determine whether or not their performance is of acceptable quality, they cannot self assess their performance and, even worse, they cannot perform well. Consequently, they will try to improve their performance by trial-and-error rather than by trying to close the gap with the criteria (Sadler, 1989). For example, if a student nurse is not aware of the importance of communicating with patients, the student may not understand why a patient is feeling uncomfortable when given an intravenous injection. In reaction to this, the student nurse might use a different needle next time rather than comforting the patient by a conversation.

Performance criteria should thus always be transparent. But for assessment to be sustainable, students should not only be able to deal with preset performance criteria, but also learn to actively seek and formulate these criteria and determine which criteria are important in judging their work, because in their future workplaces they will usually not be provided with lists of relevant criteria (Boud, 2000). For example, when a car mechanics student has to fix a motor for the first time, the student has to consider which particular criteria are important for that particular type of motor. In their study program, students thus have to learn to seek for the criteria that are important for their task performance.

Concluding this section, there are three conditions for the development of sustainable assessment skills: there is constructive alignment between instruction and assessment; students are active participants in the learning process, and performance criteria are transparent to students. The next sections take a closer look at sustainable assessment skills and their development.

Constituent Parts of Sustainable Assessment Skills

The second component is represented on the left part of the integrated model for developing sustainable assessment skills (see Figure 2.1). It pertains to the analysis of sustainable assessment skills into constituent parts. It should be noted that the teacher has a crucial role in helping students to develop sustainable assessment skills. What this role should be about is discussed in the methods for guiding students in developing sustainable assessment skills. In the current section, the ideal situation is described, being the end result of the study program. If students are taught to use assessment for sustainable aims, that is, for lifelong learning, the curriculum has to pay attention to the development of both self-assessment skills and goal-setting skills for future performance and learning. These are the two constituent parts on the second level of a so-called skills hierarchy (van Merriënboer, 1997). Based on the literature that will be discussed below, Figure 2.2 shows a more detailed hierarchy of sustainable assessment skills.

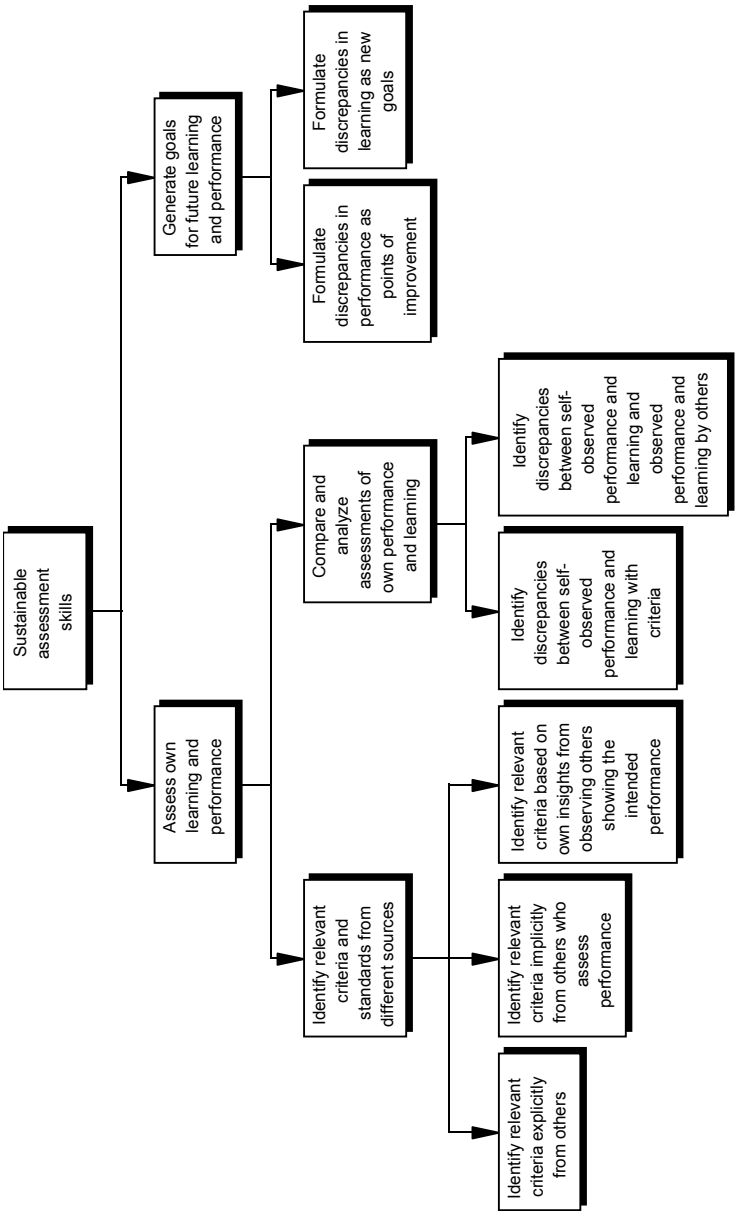

Figure 2.2: Skills hierarchy of sustainable assessment skills.

Self assess performance and learning

In order to become sustainable assessors, students must learn to assess their own performance on both single tasks (How well do I perform?) and over a series of tasks (Does my performance improve over tasks?). Performance over tasks may either increase or decrease, thereby indicating learning or a lack of learning. Assessment of one's own performance and learning requires (a) skills to identify relevant criteria from different sources (Boud, 1995), and (b) skills to critically compare and analyze different assessments of one's own performance and/or learning (see left part of the hierarchy in Figure 2.2). These are the constituent parts on the third level in this figure. Concerning performance (i.e., single tasks), students judge whether the identified criteria are reached for a particular task; concerning learning (i.e., over tasks), students judge if there is sufficient improvement of performance in the direction of the identified criteria.

Going one step further down in the hierarchy (see bottom left part of Figure 2.2), three parts constitute the ability to identify relevant criteria for performance and learning. Students must be able to identify criteria (a) explicitly provided by others, (b) implicitly used by others who assess their performance, and (c) implicitly used by others who demonstrate intended performance.

First, in school settings students often receive explicit assessment criteria, although after graduation they are usually not given such preset and explicit criteria (Tan, 2007). For example, when a student nurse for the first time enters a nursing home in which patients with dementia are treated, the student has to identify the relevant criteria for effective communication with this group of patients. Such criteria can be made explicit by a teacher, instructional materials, an internship supervisor, or a peer student. They are mostly formulated on the level of task performance. For students it is first of all important to fully understand the criteria that are explicitly provided to them, and then use these criteria to monitor if their performance over tasks is moving towards the criteria, that is, if sufficient learning is taking place.
Second, students must be able to identify the criteria, implicitly applied by others that assess their performance or learning. Teachers, peers, and relevant others (e.g., clients) all use individual implicit criteria in assessing student performance, and it is important for students to (try to) identify these criteria and invite different stakeholders to explicate them. For example, a student nurse can ask how the patient likes to be washed so the student knows what the patient expects, that is, which criteria are applied by the patient. Implicit criteria from others can be related to particular tasks but also to performance improvement over tasks. For example, a supervisor may apply specific criteria for monitoring the learning process and expect a certain degree of growth over learning tasks without explicitly stating these expectations. In such cases, it is important for students to possess the necessary skills to stimulate assessors to make their criteria explicit.

Third, students may use their observations of the performance of others to generate criteria. This is closely related to modelling, where students imitate experts or teachers who show the required performance. But this process is complex and often needs instructional support and guidance (O'Donnovan, Price, & Rust, 2004), because learners need to abstract away from concrete performances and use their insights to induce the performance criteria used by advanced task performers. For example, when students observe a supervisor or senior colleague demonstrating required performance, they should be able to deduce the important criteria from the observed performance.

One level up in the hierarchy, it is indicated that students must be able to compare and analyze different assessments (i.e., self-assessments and assessments by others) to properly assess their performance and learning (see Figure 2.2). In order to do so, they should be able to identify discrepancies (a) between their own performance and learning and previously identified criteria, and (b) between self-assessments of their performance and learning and assessments of their performance and learning by others. First, in analyzing discrepancies between their own performance and learning and criteria, students have to look at their own performance and learning and critically compare it with criteria identified from different sources. By doing so, they learn to form judgments of their own performance and learning, that is, they develop self-assessment skills (Boud, 2007). These self-assessment skills are the basis for the next step of comparing self-assessments with assessments by others.

Second, with regard to analysis of discrepancies between self-assessments (internal information) and assessments by others (external information from teachers, peers, clients, and so forth), assessments by others may be consistent or conflicting with each other and also with students' self-assessments (Paris & Paris, 2001; Sergeant, Mann, van der Vleuten, & Metsemakers, 2008). For example, a car mechanics student may think that the verbal communication with the clients is going well, while a senior car mechanic may come to an opposite assessment. Then, the student has to combine these conflicting views. A considerable number of studies have addressed students' reactions to external feedback (Black & Wiliam, 1998; Butler & Winne, 1995; Hattie & Timperley, 2007; Mory, 2003; Kluger & DeNisi, 1996). For sustainable assessment, it is important that students actively process and use the feedback they receive from other parties, and also compare it critically with their own assessments of performance and learning (Boud, 2000; Nicol & Macfarlane-Dick, 2006). Thus, the car mechanics student must acknowledge the feedback from the senior car mechanic and reflect on his own task performance to examine whether the feedback is understood. Recent studies by Eva and Regehr (2008) and Sergeant et al. (2008) confirm this idea, introducing new terms such as 'self-directed assessment seeking' and 'informed self-assessment', and stressing

that students must learn to compare external assessments and self-assessments in a reflective process.

Generate goals for future performance and learning

Based on a profound assessment of their actual performance and its development over tasks, students have to generate goals for future performance and plan their learning to reach these goals (see branch on the right in Figure 2.2). The two constituent parts lower in the hierarchy indicate that students must be able to (a) re/formulate identified discrepancies between actual and desired performance as points of improvement, and (b) set new goals for learning (Butler & Winne, 1995).

First, the formulation of points of improvement is based on discrepancies identified earlier by comparing the students' own viewpoints and the feedback provided by others (Ertmer & Newby, 1996) and should be done at a specific and observable level so that students know precisely what to work on. This will help them focus their attention on weak aspects of performance in future tasks, or alternatively, select subsequent learning tasks that are best suited to work on the identified points of improvement (Kicken et al., 2009a). For example, the student nurse who wants to improve his communication skills with demented patients first has to formulate concrete and specific points of improvement (e.g., 'always ask for confirmation to check whether the patient understands me' rather than 'improve my communication skills') and then has to work on improving exactly those aspects of the skill in subsequent tasks.

Second, apart from being able to formulate points of improvement, students should be able to set new learning goals. As stated earlier, after graduation students have to drive their own learning and thus set their own goals. Formulating learning goals is an important step in deciding what to learn and how to work towards that goal. When students have diagnosed their learning needs by exploring discrepancies, they must formulate goals and decide which learning activities can help them to fulfill these goals (Ertmer & Newby, 1996; Knowles, 1975; Loyens, Magda, & Rikers, 2008). As advocated by Eva and Regehr (2005), goals should be appropriately challenging based on earlier assessments, and for this reason earlier assessments must also be a realistic representation of students' performance levels (Eva & Regehr, 2008).

Methods for Guiding Students in Developing Sustainable Assessment Skills

The previous sections discussed, in order, the conditions for the development of sustainable assessment skills and the constituent parts of these skills. This section discusses the third component of the model (see right part of Figure 2.1), giving an overview of how students can best be guided in the development of sustainable assessment skills.

By definition, a novice in a domain is unable to make refined judgments about high-quality performance in that domain (Sadler, 1989). In the beginning of their study, students are not able to fulfill all assessment duties themselves. It is therefore important that during a study program, as domain expertise increases, assessment responsibilities are gradually handed over from the teacher to the students by providing them authentic but increasingly less guided assessment experiences. The cognitive apprenticeship model proposed by Collins, Brown, and Newman (1987; Brown, Collins, & Duguid, 1989) states that learning complex cognitive skills has to take place in authentic contexts for the learned skills to transfer outside the learning setting. The cognitive apprenticeship model presents three methods to accomplish the gradual shift of responsibilities from teacher to students: modelling, coaching, and scaffolding. First, these methods will be discussed in general, and in the following sections, these methods will be applied to the development of sustainable assessment skills. The first method, modelling, indicates that students observe the task performance of experts to learn what good performance is (i.e., performance that meets relevant criteria). The second method, coaching, indicates that observation of student performance is always followed by the provision of informative feedback, aimed at the improvement of students' task performance. The last method, scaffolding, indicates that teacher support gradually diminishes until students are able to perform the tasks independently. Together, these methods are expected to offer students the necessary guidance for developing sustainable assessment skills. The teacher has a crucial role in guiding students through this process. Table 2.2 links the two main skills for sustainable assessment, self-assessment of performance and learning and generating goals for future performance and learning, to the methods of modelling, coaching, and scaffolding. The methods will be discussed in more detail below.

Table 2.2: Sustainable assessment skills and examples of guidance for their development.

	How to guide students in the development of sustainable assessment skills?		
	Modelling	**Coaching**	**Scaffolding**
Self-assessment of performance and learning	Give modelling examples of good performance.	Provide feedback on assessment practices through in-depth discussion of criteria.	Work from given, relevant criteria to self-selected criteria; decrease informative feedback.
	Give modelling examples of good assessment.	Provide feedback on performance and learning, self-assessments, and compare self-assessments and assessments by others. Use a development portfolio to provide feedback on the learning of assessment skills over tasks	Work from assessment of students, via peer assessment, to self-assessment.
Generating goals for future performance and learning	Give modelling exam-ples of generating points of improvement and planning of learning activities.	Provide feedback on formulated learning needs and points of improvement and feed forward on planned learning activities.	Work from given points of improvement and action plans to self-formulated points of improvement and plans; decrease feedback and feed forward.

Self-assessment of performance and learning

When students leave school and enter the workplace, they must possess skills to identify relevant performance criteria and skills to compare and analyze different assessments of their performance and learning. These skills are essential for successful lifelong learners (Boud & Falchikov, 2007), and will be discussed separately.

With regard to the identification of relevant performance criteria, it is not useful to ask students to do this right from the beginning of a study program, because it can easily lead students to focus on irrelevant aspects of a task (Orsmond, Merry, & Reiling, 2002). If students need to learn to identify relevant criteria, teachers should gradually hand over this task to their students.

A first activity in teaching students to identify relevant criteria is to provide them with examples of expert performance (i.e., modelling). These examples can take the form of modelling examples which demonstrate excellent performance to students (van Merriënboer, Jelsma, & Paas, 1992). In order to be effective, these examples need to offer insight into the end result and into the entire process of reaching it, that is, the examples should be process-oriented and explain why particular performances yield desired results (van Gog, Paas, & van Merriënboer, 2008). Orsmond et al. (2002) found in an empirical study with first year Environmental Sciences and Applied Biology undergraduates that the use of exemplars in producing a poster can help students to understand the marking criteria and standards, leads to higher learning outcomes, and helps to provide meaningful feed-

back. However, the exemplars used in this study were only final products and were not process-oriented. An example of a process-oriented exemplar is to show the student how to produce the poster from scratch to end product with an explanation of each step taken. Future research is necessary to provide further empirical support for the effects of process-oriented worked examples.

A second activity in teaching students to identify relevant criteria is giving them feedback on the criteria they use (i.e., coaching). After they have been shown clear definitions of criteria along with illustrative examples as described above, students should be given the opportunity to acquire the skills for properly using the criteria. This can be achieved by offering assessment exercises where students use the presented criteria to judge other students' work and/or their own work (Nicol & Macfarlane-Dick, 2006). An alternative is described by Rust, Price, and O'Donovan (2003), who demonstrated in an empirical study with Business students the effectiveness of discussions between students and teachers about the criteria of good work, which improved their performance but not their self-assessment skills (see also Nicol & Macfarlane-Dick, 2006). The findings of Orsmond, Merry, and Reiling (2000) revealed that a discussion between students and teachers may lead to a more positive appreciation of the criteria, but not to a better understanding of the criteria as such or the ability to actually use them for the assessment of performance. So far, research results do not allow for univocal conclusions and more research on how to provide effective feedback on the identification of criteria is needed.

A final activity in teaching students to identify relevant criteria is reducing the available support (i.e., scaffolding). Early in the learning process, students should receive sufficient practice with given, relevant criteria to fine-tune the skill of identifying relevant criteria (Sadler, 1989). Practice must be supplemented by discussion and dialogue to provide informative feedback on the setting and use of criteria. Later in the learning process, learners must independently select criteria that are relevant for particular tasks from all available criteria and the provision of feedback on applied criteria gradually diminishes as the curriculum moves forward.

Modelling, coaching and scaffolding are also useful methods to guide students in the development of their skills to analyze and compare self-assessments of their performance and learning. A first activity is to offer students modelling examples of well-conducted assessments, where the teacher or an expert acts as a model when assessing task performance (i.e., modelling; van Merriënboer & Sluijsmans, 2009). Modelling examples encourage students to develop ideas of how good assessment is performed. In this process, it is necessary that teachers clearly describe the steps they take in conducting the assessment, and also explain why they take these particular steps (i.e., process-oriented modelling examples), so that students gain insight into the assessment process and the relevant considerations for reaching a

particular judgment. No research so far has dealt with this type of modelling assessments.

A second activity pertains to the provision of feedback by teachers and peers against which students can evaluate self-assessments of their performance (i.e., coaching; Nicol & Macfarlane-Dick, 2006). Taras (2003) demonstrated in an empirical study that Specialist Language students indeed learn most from self-assessments when they are combined with tutor feedback, because this helps students to be aware of all their errors. Tutor feedback should not be limited to external feedback on students' performance and learning; feedback should also be provided on the quality of self-assessments (Ertmer & Newby, 1996; Hattie & Timperley, 2007) and the process of comparing and analysing internal feedback (i.e., self-assessments) with external feedback (assessments from others). Reflective dialogue with students will help them to make optimal use of the feedback (Nicol & Macfarlane-Dick, 2006). It helps them to make better use of external feedback, to make better comparisons between internal and external feedback, and, eventually, to become better self-assessors (Taras, 2003). A development portfolio can be used to collect all received feedback (internal and external) and can make it easier for students to assess their learning over a series of tasks (Tillema et al., 2000). A coach or supervisor can help students to interpret the results of the different assessments in the portfolio and to guide the future learning process as is shown in a study with Medical students (Driessen, van Tartwijk, Overeem, Vermunt, & van der Vleuten, 2005).

A third activity to guide students in self assessing their performance and learning is by diminishing support, that is, giving students increasingly more control over their learning process (i.e., scaffolding; Corbalan, Kester, & van Merriënboer, 2008). An important condition is that students must be made aware that they are not just temporary learners, but active and future lifelong learners who must take over responsibility from their teachers (Boud & Falchikov, 2007). A gradual shift of responsibility from teacher to student has to take place, for example, by working from assessing student performance, to asking students to assess the performance of peers, to asking students to self assess their performance. Peer assessment is a powerful intermediate stage to support students in gaining self-assessment skills, because the work of fellow students is easily accessible and of the same type as their own work (Sluijsmans, Brand-Gruwel, van Merriënboer, & Martens, 2004). It is less threatening for students to assess other students' work than their own (Sadler, 1989; van Gennip, Segers, & Tillema, 2009), and the skills for peer assessment and self-assessment are comparable (Liu & Carless, 2006). In addition, students can also learn from comparing their own assessments with the assessments they receive from their peers. As van Gennip et al. (2009) describe in their review study, several conditions are necessary to enable students to work on their peer assessment skills. This relates to interpersonal features such as psychological safety and trust, and to

structural features such as peer interaction and group composition. However, more research is needed to explore these features in detail.

Generate goals for future performance and learning

Students are not used to explicitly formulating or thinking about their learning needs (Holme & Chalauisaeng, 2006). The ability to formulate learning goals presupposes the ability to formulate learning needs (i.e., discrepancies between actual and desired performance), related points of improvement, and learning activities. Therefore, it is important that students are explicitly encouraged to think about ways of improving their own performance without possibly negative consequences for their results. The skills of generating goals of future performance and future learning will be discussed separately.

The use of modelling examples (i.e., modelling) is a first activity to help students develop the skill of formulating points of improvement. This can be done by a teacher or coach who demonstrates the interpretation of assessment results and the formulation of points of improvement as is shown in a study by Kicken, Brand-Gruwel, van Merriënboer, and Slot (2009b). In their research in a Hairdressing program in secondary vocational education, they used a development portfolio which, besides a self-assessment instrument, focused on the formulation of points of improvement and task selection aspects. They showed that with the necessary teacher support and advice, a development portfolio leads to the formulation of better points of improvement.

As a next activity, students should be given feedback by their teachers or supervisors on the points of improvement students have formulated on the basis of assessments from others and/or self-assessments (i.e., coaching). Once students are aware of the real causes of their shortcomings they are better able to formulate points of improvement. Assessors can also formulate points of improvement - or students can ask them to do so - enabling students to compare and contrast their own points of improvement with those suggested by others (Kicken et al., 2009b).

As a final activity, the support provided by the teacher or coach should diminish as students gain more experience in formulating points of improvement (i.e., scaffolding). For example, first the teacher may formulate points of improvement for a particular student based on available assessment results, then the student may formulate points of improvement with feedback from the teacher, and finally the student formulates the points of improvement without any support. No research so far has shed light on this diminishing amount of support in formulating points of improvement.

Modelling examples (i.e., modelling) can also be used for formulating new learning goals over several learning tasks. According to Kicken et al. (2009b), this can be done by a teacher or coach who explains why particular learning activities are chosen to meet the identified goals, based on the points of improvement. In the

same study it appeared that when students got advice in combination with a development portfolio, they were better able to choose future learning activities to work on.

As a next activity, students should be given feedback by their teachers or supervisors on the learning goals they formulate based on the points of improvement formulated earlier (i.e., coaching). In addition, coaching should not be limited to feedback, in which students look back with their teachers to explore weaknesses and the causes thereof, but also include feed-forward, in which advice is given on future learning activities to resolve the weaknesses. Detailed points of improvement make it easier to identify relevant learning activities as was shown in the research of Kicken et al. (2009b). Feedback and feed-forward may pertain to both the level of single-task performance and the level of multiple-task learning. Together with a coach or supervisor, students can look at their learning progress as documented in a development portfolio and formulate learning needs, points of improvement, and new learning activities based on that overview (Driessen et al., 2005).

As a final activity, the support provided by the teacher should diminish as students gain more experience in planning future learning activities and setting new learning goals (i.e., scaffolding). For example, first the teacher may formulate future learning activities for a particular student based on the points of improvement formulated earlier, then the student may formulate new learning goals with feedback from the teacher or coach, and finally the student plans his or her future learning activities independently. No research so far has shed light on this diminishing amount of support in planning future learning activities and setting new learning goals.

Discussion

This chapter discussed an integrated model for the development of sustainable assessment skills. We first compared the features of sustainable assessment with the features of summative and formative assessment. Past research offered valuable insights into organizing effective and efficient summative and formative assessments, but our knowledge regarding the use of sustainable assessment is still in its infancy.

The proposed model describes an ideal situation and consists of three components: (a) conditions that must be fulfilled for sustainable assessment skills to develop, (b) constituent parts of sustainable assessment skills, and (c) methods for the development of sustainable assessment skills. As for conditions, there must be constructive alignment between instruction and sustainable assessment; students should be active learners, and performance criteria must be transparent to the learners. With regard to constituent parts of sustainable assessment, the skill to self

assess performance and learning, the skill to generate goals for future performance and learning, and several lower-level enabling skills were identified. With regard to methods for teaching sustainable assessment skills, cognitive apprenticeship and its related methods of modelling, coaching, and scaffolding offer valuable suggestions for instruction.

Since this model is a first attempt at providing a framework for the concept of sustainable assessment and its development, some critical remarks should be made. First, the presented model is based on a literature review including more conceptual than empirical papers, making it impossible to draw definite conclusions. Further empirical research is needed to test whether the proposed model needs adaptations and/or extensions. It is likely that future research will increase our understanding of additional conditions that need to be fulfilled and will provide more detailed insights that can be included in the skills hierarchy. For example, additional conditions may pertain to the use of tools (e.g., portfolios) or the training of teachers in guiding students, and the planning of future learning may need a more detailed description of the skills hierarchy. The cognitive apprenticeship model of Collins et al. (1987) was taken as a sound starting point for the description of methods to guide students in their development of sustainable assessment skills, but possibly other models can be used to come up with alternative methods.

There are some important research implications that can be derived from the model. First, future research might focus on the provision of high-standard exemplars in the form of modelling examples, where teachers demonstrate how to perform according to relevant criteria, how to assess on the basis of these criteria, and how to formulate points of improvement and plan future learning. Process-oriented examples pay explicit attention to why particular actions are taken, and their use could be coupled with research on the effects of group discussions with students. Second, research might focus on the effects of feedback and coaching on the different constituent parts of sustainable assessment, including the (in)correct use of performance criteria, the quality of self-assessments, comparisons of self-assessments and assessments by others, the formulation of learning needs and points of improvement, and the formulation of new learning goals. Special attention could be devoted to the role of development portfolios in providing feedback on the development of sustainable assessment skills over a series of tasks, that is, on the development of assessment skills throughout the educational program. Third, more research is needed to clarify how scaffolding of all sustainable assessment skills can best be organized.

A practical implication that follows from the model is that teachers and policy makers in professional and vocational training should consider the development of sustainable assessment skills as a valuable goal of educational programs. They should ensure that the conditions for the development of sustainable assessment skills are met, and also stimulate educational innovations that aim to develop sus-

tainable assessment skills. Some of the methods discussed in this chapter have already proven to work in educational practice and could be implemented in schools. For example, the research of Kicken et al. (2009b) showed that the use of a development portfolio in combination with regular coaching settings can help students to make more use of assessments, plan their future learning more deliberately, and gain more insight into their learning process.

In conclusion, sustainable assessment can be regarded as a promising idea to help students to become lifelong learners, but further development is needed to determine how sustainable assessment skills can best be taught. This chapter offered some starting points for a theoretical foundation of sustainable assessment and gave suggestions for future research on this topic. It is noteworthy to mention that such research should not be restricted to in-school educational programs but also be extended to include graduates' workplace learning. Only then will it become clear if the guidelines proposed in this chapter can actually contribute to the development and application of sustainable assessment skills.

References

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, 56*, 523-538.

Biggs, J. (1996). Enhancing teaching through constructive alignment. *Higher Education, 32,* 347-364.

Black, P., & Wiliam, D. (1998). Assessment and classroom learning. *Assessment in Education: Principles, Policy and Practice, 5*, 7-74.

Boekaerts, M., Pintrich, P. R., & Zeidner, M. (2000). *Handbook of self-regulation*. San Diego: Academic Press.

Boud, D. (1995). *Enhancing learning through self-assessment.* London: Kogan Page.

Boud, D. (2000). Sustainable assessment: Rethinking assessment for the learning society. *Studies in Continuing Education, 22*, 151-167.

Boud, D. (2007). Reframing assessment as if learning were important. In D. Boud & N. Falchikov (Eds.), *Rethinking assessment in higher education: Learning for the longer term* (pp. 14-25). New York: Routledge.

Boud, D., & Falchikov, N. (2006). Aligning assessment with long-term learning. *Assessment and Evaluation in Higher Education, 31*, 399-413.

Boud, D., & Falchikov, N. (2007). Developing assessment for informing judgement. In D. Boud & N. Falchikov (Eds.), *Rethinking assessment in higher education: Learning for the longer term* (pp. 181-197). New York: Routledge.

Brown, J. S., Collins, A., & Duguid, P. (1989). Situated cognition and the culture of learning. *Educational Researcher, 18*, 32-42.

Butler, D. L., & Winne, P. H. (1995). Feedback and self-regulated learning: A theoretical synthesis. *Review of Educational Research, 65*, 245-281.

Carless, D. (2007). Learning-oriented assessment: Conceptual bases and practical implications. *Innovations in Education and Teaching International, 44*, 57-66.

Corbalan, G., Kester, L., & van Merriënboer J. J. G. (2008). Selecting learning tasks: Effects of adaptation and shared control on learning efficiency and task involvement. *Contemporary Educational Psychology, 33,* 733-756.

Collins, A., Brown, J. S., & Newman, S. E. (1987). *Cognitive apprenticeship: Teaching the craft of reading, writing, and mathematics* (Tech. Rep. No. 403). BNN Laboratories, Cambridge, MA. Centre for the Study of Reading, University of Illinois.

Dunn, K. E., & Mulvenon, S. W. (2009). A critical review of research on formative assessments: The limited scientific evidence of the impact of formative assessments in education. *Practical Assessment, Research, and Evaluation, 14*(7).

Dochy, F., Segers, M., & Sluijsmans, D. (1999). The use of self-, peer and co-assessment in higher education: A review. *Studies in Higher Education, 24*, 331-350.

Driessen, E. W., van Tartwijk, J., Overeem, K., Vermunt, J. D., & van der Vleuten, C. P. M. (2005). Conditions for successful reflective use of portfolios in undergraduate medical education. *Medical Education, 39*, 1230-1235.

Ertmer, P. A., & Newby, T. J. (1996). The expert learner: Strategic, self-regulated, and reflective. *Instructional Science, 24,* 1-24.

Eva, K. W., & Regehr, G. (2005). Self-assessment in the health professions: A reformulation and research agenda. *Academic Medicine, 80*, S46-S54.

Eva, K. W., & Regehr, G. (2008). ´´I'll never play professional football'' and other fallacies of self-assessment. *The Journal of Continuing Education in the Health Professions, 28*, 14-19.

Gibbs, G., & Simpson, C. (2004). Conditions under which assessment supports student learning. *Learning and Teaching in Higher Education, 1*, 3-31.

Govaerts, M. J. B., van der Vleuten, C. P. M., Schuwirth, L. W. T., & Muijtjens, A. M. M. (2007). Broadening perspectives on clinical performance assessment: Rethinking the nature of in-training assessment. *Advances in Health Sciences Education, 12,* 239-260.

Gulikers, J., Bastiaens, T., & Kirschner, P. A. (2004). A five-dimensional framework for authentic assessment. *Educational Technology Research and Development, 52*, 67-85.

Hattie, J., & Timperley, H. (2007). The power of feedback. *Review of Educational Research, 77*, 81-112.

Holme, R., & Chalauisaeng, B. (2006). The learner as needs analyst: The use of participatory appraisal in the EAP reading classroom. *English for Specific Purposes, 25,* 403-419.

Kerka, S. (1995). *Techniques for authentic assessment: Practice application brief*. Columbus: ERIC Clearinghouse on Adult, Career, and Vocational Education

Kicken, W., Brand-Gruwel, S., van Merriënboer, J. J. G., & Slot, W. (2009a). Design and evaluation of a development portfolio: How to improve students' self-directed learning skills. *Instructional Science, 37*, 453-473.

Kicken, W., Brand-Gruwel, S., van Merriënboer, J. J. G., & Slot, W. (2009b). The effects of portfolio-based advice on the development of self-directed learning skills in secondary vocational education. *Educational Technology Research and Development, 57,* 439-460.

Kluger, A. N., & DeNisi, A. (1996). The effects of feedback interventions on performance: Historical review, a meta-analysis and a preliminary feedback intervention theory. *Psychological Bulletin, 119,* 254-284.

Knight, P. T. (2002). Summative assessment in higher education: Practices in disarray. *Studies in Higher Education, 27*, 175-286.

Knowles, M. (1975). *Self-directed learning: A guide for learners and teachers.* Englewood Cliffs, NJ: Prentice Hall/Cambridge.

Liu, N., & Carless, D. (2006). Peer feedback: The learning element of peer assessment. *Teaching in Higher Education, 11*, 279-290.

Loyens, S. M. M., Magda, J., & Rikers, R. M. J. P. (2008). Self-directed learning in problem-based learning and its relationships with self-regulated learning. *Educational Psychology Review, 20,* 411-427.

McDonald, R., Boud, D., Francis, J., & Gonczi, A. (1995). *New perspectives on assessment* (pp. 45). Paris: UNESCO.

Mory, E. H. (2003). Feedback research revisited. In D. H. Jonassen (Ed.), *Handbook of Research on Educational Communications and Technology* (pp. 745–783). New York: MacMillian.

Nicol, D. J., & Macfarlane-Dick, D. (2006). Formative assessment and self-regulated learning: A model and seven principles of good feedback practice. *Studies in Higher Education, 31*, 199-218.

Nieweg, M. R. (2004). Case study: Innovative assessment and curriculum redesign. *Assessment and Evaluation in Higher Education, 29*, 203-214.

O'Donnovan, B., Price, M., & Rust, C. (2004). Know what I mean? Enhancing student understanding of assessment standards and criteria. *Teaching in Higher Education, 9,* 325-335.

Orsmond, P., Merry, S., & Reiling, K. (2000). The use of student derived marking criteria in peer and self-assessment. *Assessment and Evaluation in Higher Education, 25*, 23-38.

Orsmond, P., Merry, S., & Reiling, K. (2002). The use of exemplars and formative feedback when using student derived marking criteria in peer and self-assessment. *Assessment and Evaluation in Higher Education, 27*, 309-323.

Paris, S. G., & Paris, A. H. (2001). Classroom applications of research on self-regulated learning. *Educational Psychologist, 36*, 89-101.

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

Rust, C., Price, M., & O'Donovan, B. (2003). Improving students' learning by developing their understanding of assessment criteria and processes. *Assessment and Evaluation in Higher Education, 28*, 147-164.

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

Sadler, D. R. (1998). Formative assessment: Revisiting the territory. *Assessment in Education, 5*, 77-84.

Sadler, D. R. (2005). Interpretations of criteria-based assessment and grading in higher education. *Assessment and Evaluation in Higher Education, 30*, 175-194.

Sergeant, J., Mann, K., van der Vleuten, C., & Metsemakers, J. (2008). ''Directed'' self-assessment: Practice and feedback within a social context. *Journal of Continuing Education in the Health Professions, 28*, 47-54.

Shute, V. J. (2008). Focus on formative feedback. *Review of Educational Research, 78*, 153-189.

Sluijsmans, D. M. A., Brand-Gruwel, S., van Merriënboer, J. J. G., & Martens, R. (2004). Training teachers in peer-assessment skills: Effects on performance and perceptions. *Innovations in Education and Teaching International, 41*, 59-78.

Sluijsmans, D. M. A., Straetmans, G., & van Merriënboer, J. J. G. (2008). Integrating authentic assessment with competency based learning: The Protocol Portfolio Scoring. *Journal of Vocational Education and Training, 60,* 157-172.

Tan, K. H. K. (2007). Conceptions of self-assessment: What is needed for long-term learning. In D. Boud & N. Falchikov (Eds.), *Rethinking assessment in higher education: Learning for the longer term* (pp. 181-197). New York: Routledge.

Tan, K. H. K. (2008). Qualitatively different ways of experiencing student self-assessment. *Higher Education Research and Development, 27*, 15-29.

Taras, M. (2003). To feedback or not to feedback in student self-assessment. *Assessment and Evaluation in Higher Education, 28*, 549-565.

Taras, M. (2005) Assessment - summative and formative - some theoretical reflections. *British Journal of Educational Studies, 35*, 466-478.

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

Van Gennip, N. A. E., Segers, M. S. R., & Tillema, H. H. (2009). Peer assessment for learning from a social perspective: the influence of interpersonal variables and structural features. *Educational Research Review, 4*, 41-54.

Van Gog, T., Paas, F., & van Merriënboer, J. J. G. (2008). Effects of studying sequences of process-oriented and product-oriented worked examples on troubleshooting transfer efficiency. *Learning and Instruction, 18*, 211-222.

Van Merriënboer, J. J. G. (1997). *Training complex cognitive skills*. Englewood Cliffs, NJ: Educational Technology Publications.

Van Merriënboer, J. J. G., Jelsma, O., & Paas, F. G. W. C. (1992). Training for reflective expertise: A four-component instructional design model for complex cognitive skills. *Educational Technology Research and Development, 40*, 23-43.

Van Merriënboer, J. J. G., & Sluijsmans, D. M. A. (2009). Toward a synthesis between cognitive load theory and self-directed learning. *Educational Psychology Review, 21*, 55-66.

CHAPTER 3

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

Abstract

In this study in secondary vocational education, students in Nursing and Care (N = 68) work on learning tasks, self assess task performance, and formulate points of improvement. In a highlight group students' attention is drawn to the assessment criteria that are relevant for a particular learning task by emphasizing them in a list with all possible criteria. It is compared to a no-highlight group in which students receive the list with all possible criteria for each task. Students in the relevant criteria group outperformed the students in the all criteria group in the test task, but they experienced a higher mental effort for conducting the assessment. Care students in the relevant criteria group were better in generating points of improvement than care students in the all criteria group. Furthermore, nursing students outperformed care students in the test task and care students selected more criteria than nursing students.

This chapter is based on: Fastré, G. M. J., van der Klink, M. R., Sluijsmans, D., and van Merriënboer, J. J. G. (2010). *Drawing Students' Attention to Relevant Assessment Criteria: Effects on Self-Assessment Skills and Performance.* Manuscript submitted for publication.

The nursing profession asks from nurses to engage in multiple tasks under a high cognitive load, often leading to stressful situations (Jourdain & Chênevert, 2010). Next to the direct patient care, nurses also have to perform indirect patient care like charting, preparing medications and coordinating care (Wolf et al., 2006). Student nurses therefore must acquire the professional competences needed to perform well in the workplace. But in addition they need to develop self-directed learning skills, such as self assessing their performance and judging own strengths and weaknesses, in order to cope with the uncertain, unpredictable and constantly changing circumstances in their future workplaces (Boud & Falchikov, 2006; Kicken, Brand-Gruwel, van Merriënboer, & Slot, 2009).

In this study, self-assessment is defined as the participation of students in the process of selecting performance criteria from a predefined set of criteria, making judgments about the extent to which relevant criteria are met during the performance of particular learning tasks, and generating points of improvement for future learning tasks. As can be seen from Figure 3.1, learning self-assessment is a cyclical process in which the outcomes of the self-assessment on previous tasks are used as inputs for future tasks (Eva & Regehr, 2005). For instance, outcomes of the self-assessment process may refer to identified points of improvement that help learners to focus their attention on particular aspects of a next learning task.

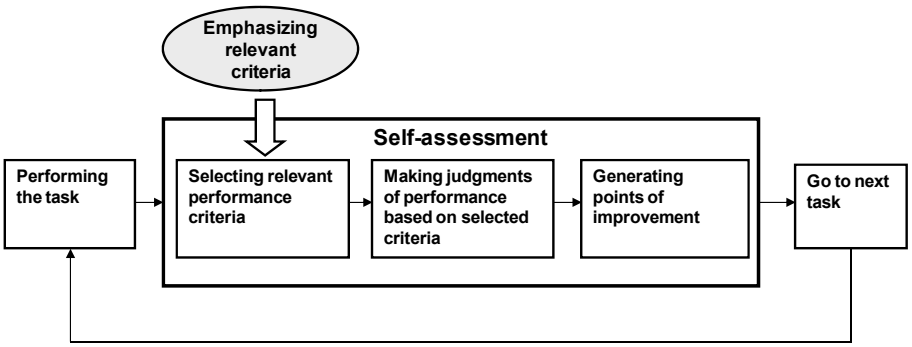

Figure 3.1: Model of the cyclical learning process with self-assessment at the heart.

As Boud (1999) suggests, students' self-assessment is often based on cues they collect from teachers and peers while working on learning tasks. Therefore, self-assessment is not a fully individual and isolated activity, but an integral part of a social learning process. The main question of the study reported in this chapter concerns: Does supporting students in selecting relevant performance criteria for assessing particular learning tasks help them to improve their task performance and self-assessment skills?

A considerable number of studies has been conducted on students' skills to assess their own performance (for reviews, see Boud & Falchikov, 1989; Dochy,

Segers, & Sluijsmans, 1999; Dunning, Heath, & Suls, 2004; Falchikov & Boud, 1989). These studies show that the relationship between students' self-assessment of performance and the true quality of their performance is rather weak. There may be at least three reasons for this. First, students usually have little or no experience with self-assessment. In their prior schooling, students have mostly been the subjects rather than the actors of assessment, they were not encouraged to take own responsibility for the assessment process, and they only took tests and received feedback on their performance (Boud & Falchikov, 2006). Second, students' judgements are often biased. They face difficulties in recognizing their own incompetence (Kruger & Dunning, 1999), and less advanced students usually overestimate their own performance (Dunning et al., 2004). Third, the way students rate their own work is often different from the way teachers rate the same work, which is reinforced by a lack of explicit performance criteria for self-assessment (Boud & Falchikov, 1989).

In the reported study, students are given the opportunity to practice self-assessment skills and receive feedback on the quality of self-assessments by comparing their self-assessments with teacher assessments. The focus of this chapter is yet on the problem that students are insufficiently able to decide which performance criteria are relevant for a particular task, that is, on the first step of the cyclical process of self-assessment (Figure 3.1; see also Orsmond, Merry, & Reiling, 2002). Two sub-problems are that students are not able to (1) independently devise appropriate criteria, and (2) select the appropriate criteria for one particular learning task, not even when they are given the whole set of possibly relevant criteria for all tasks in the domain.

With regard to the first problem, Boud and Falchikov (1989) state that the identification of performance criteria to be applied to one's work is an important and integral part of self-assessment. As part of their educational program, students need training and practice in developing the skill to devise appropriate criteria, as was stated in studies conducted by Boud and Brew (1995) and Orsmond et al. (2002). Dunning et al. (2004), however, point at considerable problems novices experience when they must formulate their own performance criteria: Their lack of domain expertise causes mistakes and also prevents them from recognizing these mistakes since they do not have an adequate view yet on what is good performance. For example, if student nurses have no experience in nursing patients, they are not able to know which competences are important in the contacts with patients, or what constitutes 'good practice' for these contacts. Therefore, it is appropriate to use pre-specified performance criteria for novice students.

Concerning the second problem, students could reasonably consider a very large set of potential performance criteria when they have to perform and assess real-life whole tasks representative of professional life (Sadler, 1989). For each performed task, this whole set of criteria can be split up into two parts: relevant criteria

and irrelevant criteria. For example, students in nursing should be able to master the entire competence of nursing with a substantial number of criteria to meet. However, for washing a highly demented male patient another subset of criteria is relevant than for feeding a twenty-year old female patient. Sadler (1989) argues that it is important for students to become competent in taking an adequate decision on which criteria are relevant, and which criteria are irrelevant, for particular tasks. When students select assessment criteria, four different situations can occur: (1) correct selection of a relevant criterion as being relevant (true positive); (2) incorrect selection of an irrelevant criterion as being relevant (false positive); (3) incorrect selection of a relevant criterion as being irrelevant (false negative), and (4) correct selection of an irrelevant criterion as being irrelevant (true negative). In the ideal situation, students select only true positives and avoid false positives.

In Dutch secondary vocational education, students are often confronted with long lists of pre-specified criteria without any information or training helping them to determine which criteria are actually relevant for the task at hand (Kicken, Brand-Gruwel, & van Merriënboer, 2008; Corbalan, Kester, & van Merriënboer, 2006). Regehr and Eva (2006) and Dunning et al. (2004) describe the risk of overloading students with too many criteria, from which they must autonomously select the relevant criteria for that task. Students will often only select the criteria for skills on which they already perform well or which they like, because people naturally strive at creating positive emotions and will therefore not easily recognize inadequacies in their performance. Consequently, students must be explicitly stimulated to learn to select relevant criteria for tasks at hand.

The process of learning to select relevant criteria may be hampered by the limited processing capacity of the human mind, because a very large number of combinations of relevant criteria can be selected from the whole set of possibly relevant criteria (van Merriënboer & Sweller, 2005). Next to that, the nursing task itself already poses a high cognitive load (Wolf et al., 2006). This load related to the nursing task is called intrinsic cognitive load (Sweller, van Merriënboer, & Paas, 1998). Thus, the process of selecting relevant criteria may be so overwhelming that students have insufficient cognitive capacity available to conduct the nursing task well and simultaneously assess their performance on those criteria. Furthermore, the load placed on students' working memory when they have to select relevant criteria can be either positive or negative (Sweller et al., 1998). It is positive if it is caused by learning processes leading to a better comprehension of the criteria (germane load), a better task performance, and a more accurate self-assessment, such as the construction of cognitive schemas that link particular features of tasks to the (ir)relevance of particular criteria. For example, when student nurses see the whole list of possibly relevant criteria they may become aware of the fact 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 placed on working memory is negative when it is caused by processes not effective for learning (extraneous load), such as incorrectly identifying irrelevant criteria as being relevant. For example, when students see the whole list of criteria they may be tempted to apply irrelevant criteria just 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 the relevant criteria. This study explores the effects of selecting criteria on perceived cognitive load as well as on task performance and the accuracy of self-assessment.

The main goal of this study is to investigate the effects of drawing students' attention to relevant assessment criteria on task performance and self-assessment skills. For one group of students the criteria that are relevant for specific tasks that must be assessed are highlighted in the whole set of criteria, whereas for another group of students no highlighting occurs (i.e., they receive an undifferentiated list with all possibly relevant criteria). The first hypothesis is that students who practice with the relevant criteria will show higher task performance on a test task than students in the all criteria group, because they know better what good task performance looks like (Dochy et al., 1999). The second hypothesis is that students who practice self-assessment skills with the relevant criteria will be better in selecting relevant criteria in a test phase than students who practice with all criteria. The third, related hypothesis is that students who practice with the relevant criteria will be better in avoiding selecting irrelevant criteria in the test phase. As an exploration, the perceived cognitive load of students who self assess performance with the relevant criteria and with all criteria will also be compared. Because selecting the relevant performance criteria is the first step in a cyclical process of learning self-assessment (see Figure 3.1), the fourth hypothesis is that students who practice with the relevant criteria will show more accurate judgements of their performance than students who practice with all criteria. Finally, to close the cyclical process of self-assessment, the fifth hypothesis is that students who practice with the relevant criteria will be better able to formulate points of improvement than the other group, because the given support in identifying relevant strengths and weaknesses helps them to set appropriate learning goals for future tasks (Eva & Regehr, 2005). Next to these hypotheses it will be explored whether there are differences between nursing and care students with regard to task performance and self-assessment skills. Nursing students are level 4 students in the European Qualifications Framework, meaning that they have to acquire factual and theoretical knowledge in broad contexts within the nursing field. Their study program lasts for 4 years. Care students on the other hand, are level 3 students, meaning that they have to acquire knowledge of facts, principles, processes, and general concepts within the nursing field. Their study program lasts for 3 years. The major difference between the pro-

grams is the more theoretical orientation of nursing students and the more practical orientation of care students. The first year of the study program is partly common for nursing and care students. 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 if they need a different kind of support in using the assessment criteria as they have a different orientation in their study program.

Method

Participants and design

First-year students attending programs in Nursing and Care in secondary vocational education (N = 68; 6 males and 62 females) participated in the experiment. These students spend part of their time at school and work part of their time in practice. The study was set up as a 2 x 2 factorial design with the between-subjects factors Relevance (relevant vs. all criteria) and Program (nursing vs. care). The data of 10 students were lost due to a technical problem with the learning environment. From the remaining students, 18 students are in the relevant criteria/nursing group, 18 students in the all criteria/nursing group; 16 students in the relevant criteria/care group, and 6 students in the all criteria/care group.

Learning Materials

The electronic learning environment 'Care Village' (see Figure 3.2) was used for data collection as part of their regular study program. It is a representation of an authentic work environment in which all relevant care and nursing settings are available (e.g., hospital, psychiatric hospital, elderly care; Gulikers, Bastiaens, Kirschner, & Kester, 2008). Furthermore, students can visit their virtual school and a multimedia centre.

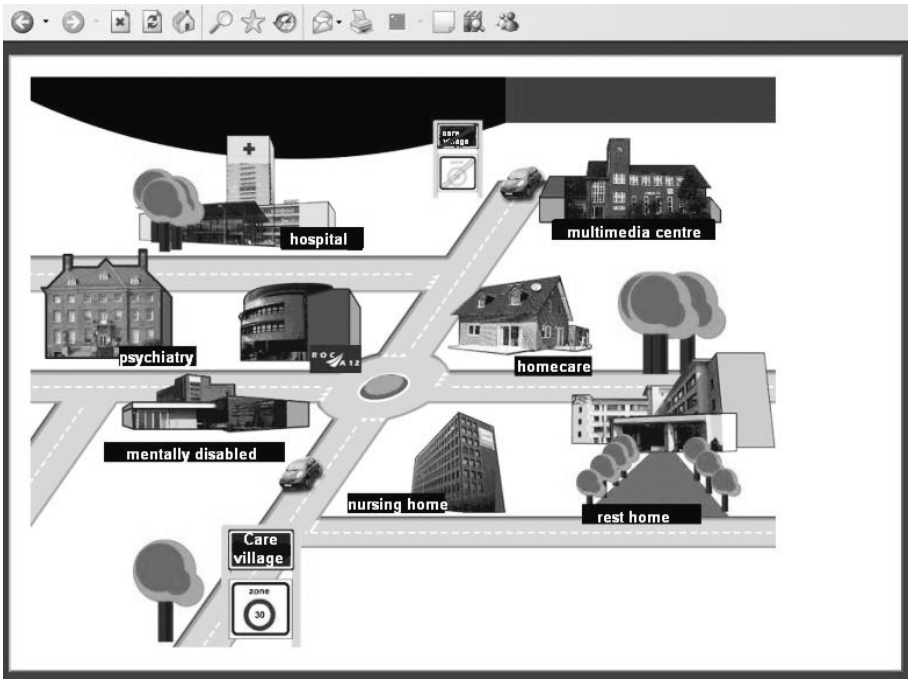

Figure 3.2: Screen dump of the homepage of Care Village.

In Care Village, students can select learning tasks from a care setting of their choice. No distinction was made for nursing and care students. 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 (4C/ID; van Merriënboer, 1997; van Merriënboer & Kirschner, 2007): (1) worked-out examples with the assignment to study a task performed by someone else and answer questions about it, (2) completion tasks with the assignment to perform only part(s) of a task, and (3) conventional tasks with the assignment to perform the whole task independently. Each learning task consisted of a case description, leading questions to understand the case, an assignment to perform at school, and a second assignment to perform during students' internships. Thus, students first read the task description in Care Village and then perform the assignments in school and at the workplace. For most learning tasks, the assignment is to nurse a simulation patient in a simulated setting (school) and in real life (workplace), and to complete a nursing dossier.

After students finish a learning task, they self assess their performance and generate points of improvement for next learning tasks. 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 entire set of tasks a

nurse has to perform at beginners' level. Figure 3.3 presents a screen dump of Care Village showing a small part of this list with assessment criteria.

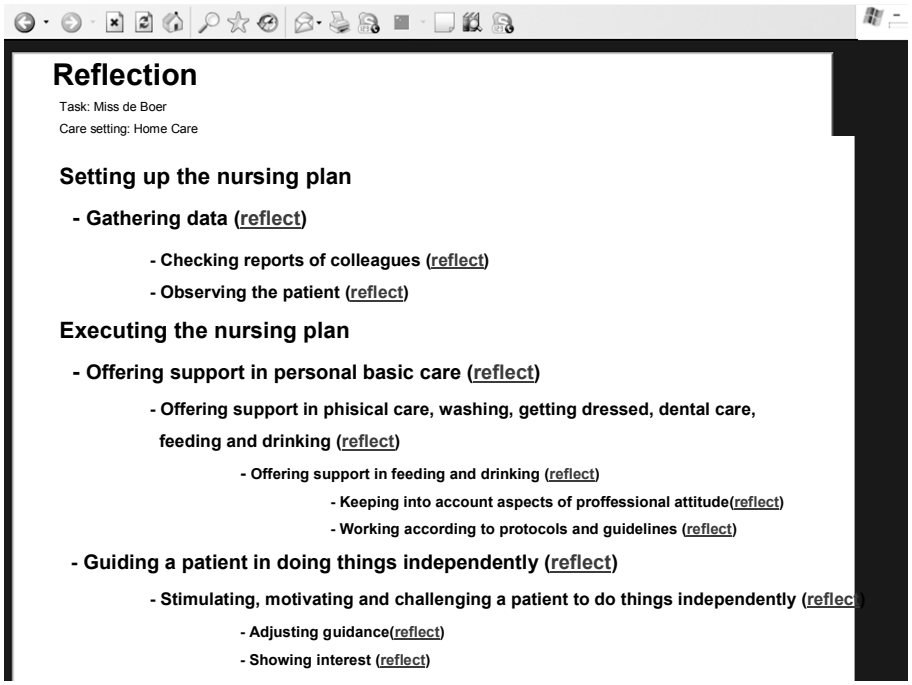

Figure 3.3: Screen dump of the list with assessment criteria in Care Village.

The assessment criteria are made operational by means of scoring rubrics in which students indicate their competence level in relation to each relevant criterion (Sluijsmans, Straetmans, & van Merriënboer, 2008). Figure 3.4 presents another screen dump of Care Village showing an example of a scoring rubric for the criterion 'conducting an anamnesis'.

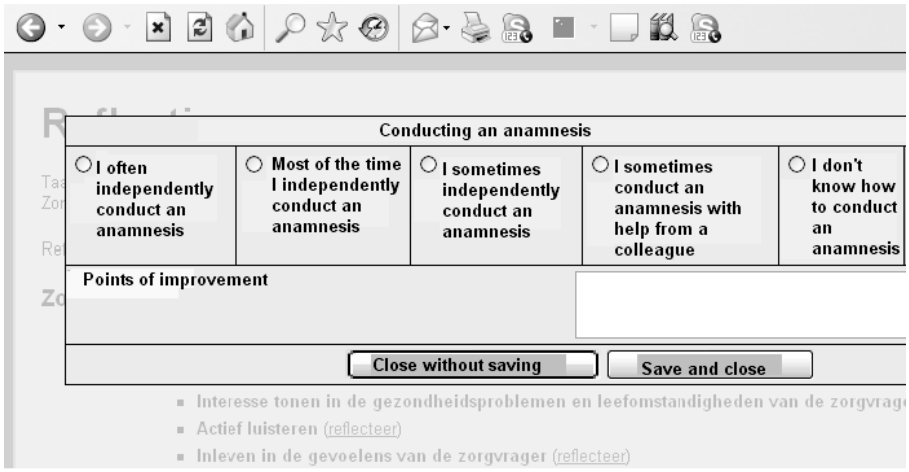

Figure 3.4: Screen dump of the scoring rubric 'conducting an anamnesis' in Care Village.

Students from all experimental groups could choose from the same set of learning tasks and received the same set of 69 assessment criteria. However, in the relevant criteria groups, the criteria that were relevant for the to-be-assessed learning task were displayed in bold and the criteria that were not relevant for the assessed task were displayed in normal print; in the all criteria groups, all criteria were displayed in bold. Consequently, students in the relevant criteria groups got a complete overview of all available criteria but were alerted that some criteria were relevant and others were not. They self assessed their performance only on the relevant criteria. Students in the all criteria groups had to select the criteria they assumed to be relevant and self assessed their performance on those self-selected criteria. All students had to self assess their performance on all learning tasks they worked on over a period of four months. Students' task performance was also assessed by the teacher, using the same list of criteria as the students. The virtual school of Care Village allows students to compare their self-assessments with the assessments of their teacher.

The final task students had to perform was a test task. This was a conventional task resembling the learning tasks. However, all students were confronted with the list of 69 assessment criteria with no criteria printed in bold. Thus, they were requested to select the criteria they assumed relevant and assess themselves on those criteria. The test task only consisted of the part to perform in school and not in the workplace. For all students, their actions in Care Village (e.g., number and nature of learning tasks performed, self-assessments, and teacher assessments) were logged automatically.

Measurements

Selection of relevant criteria

The correct selection of relevant criteria during self-assessment indicates how competent students are in selecting the criteria relevant for a particular learning task from the whole set of 69 criteria. To measure this, the criteria selected by the student were compared with the relevant criteria as indicated by the expert teachers. Selection of relevant criteria is computed as the amount of true positives selected by the student divided by the total amount of relevant criteria possible (i.e., true positives). A minimum score of 0 and a maximum score of 1 can be obtained with 1 being the most optimal score (i.e., all true positives were selected by the student).

Selection of irrelevant criteria

The avoidance of selecting irrelevant criteria during self-assessment indicates how competent students are in ignoring the criteria irrelevant for a particular learning task, so not choosing true negatives. To measure this, the criteria avoided by the student were compared with the irrelevant criteria as indicated by the expert teachers. Selection of irrelevant criteria is computed as the amount of true negatives selected by the student divided by the total amount of irrelevant criteria (i.e., true negatives). A minimum score of 0 and a maximum score of 1 can be obtained with 1 being the most optimal score (i.e., no false positives were selected by the student).

Accuracy of self-assessment

The accuracy of self-assessment is based on the degree of agreement between the self-assessment of the student and the teacher assessment. 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. Rust, Price, and O'Donovan (2003) described a simple numerical system to compare scores of teachers and students. In their system, 0 indicates that student and teacher selected the same score; 1 indicates a one-grade difference (+1 if the student's grade is higher, -1 if the student's grade is lower than that of the teacher), 2 indicates a two-grade difference, et cetera. For our purpose, the absolute value of the difference was taken as a measure for accuracy of self-assessment, with 0 being the optimal score.

Generation of points of improvement

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

Learning task performance
Learning task performance indicates how well a student executed a learning task. It is based on the teacher's mean assessment score over all relevant criteria for the task, using a 4-point rating scale for each criterion, thus yielding a maximum score of 4.

Test task performance
Test task performance is measured in the same way as learning task performance, using the same 4-point rating scale. Two independent raters different from the teacher assessed task performance. Interrater reliability was high, Pearson r = .92, p = .01.

Self-assessment mental effort
After self-assessments of the learning tasks and test task, students were asked to rate their mental effort required to perform the self-assessment with a 7-point cognitive-load rating scale as used in an experiment in secondary vocational education by Corbalan, Kester, and van Merriënboer (2008), ranging from a very small amount of mental effort (1) to a very high amount of mental effort (7).

Student perceptions
Student perceptions of the following seven aspects were measured with a 4-point Likert scale: relevance of self-assessment, ability to self assess, interesting course material, task orientation, interest and pleasure in the learning tasks, interest and pleasure in reflection, and usefulness. The self-directed learning skills questionnaire adapted from 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 on the learning environment. Three scales of the Intrinsic Motivation Inventory by Deci, Eghrari, Patrick, and Leone (1994), translated into Dutch by Martens and Kirschner (2004), were used to measure interest and pleasure in the learning tasks, interest and pleasure in reflection, and usefulness of the learning environment. The Cronbach's alpha scores of the perception scales ranged from .66 to .93 for the population under study and are thus acceptable to high.

Procedure

At the start of the experiment, students and teachers received written and verbal instructions on how to work in the learning environment Care Village. During a period of four months (learning phase), students were requested to work on the learn-

ing tasks, to self assess their performance on each task, and to rate the mental effort experienced during self-assessment. At the end of the experiment, students were requested to work on the test task for which the self-assessment and mental effort rating was the same as for the learning tasks (test phase). Finally, students were asked to fill out the seven student perception scales.

Results

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

Table 3.1 presents the means and standard deviations for the dependent variables in the test and learning phase.

Table 3.1: 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

Test Phase

On test task performance, a main effect of Relevance was found, $F(1, 54) = 3.750$, *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). Furthermore, a main effect of Program was found, $F(1, 54) = 7.435$, *MSE* = 3.747, *p* = .009, $\eta_p^2 = .121$, indicating that 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.

On selection of relevant criteria, a main effect of Program was found, $F(1, 54) = 5.407$, *MSE* = .561, *p* = .024, $\eta_p^2 = .091$, 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). Neither effect of Relevance nor an interaction effect was found on selection of relevant criteria.

On selection of irrelevant criteria, a main effect of Program was found, $F(1, 54) = 7.068$, *MSE* = .604, *p* = .01, $\eta_p^2 = .116$, 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). Neither effect of Relevance nor an interaction effect was found on selection of relevant criteria.

On accuracy of self-assessment, no significant effects were found.

On generation of points of improvement, no main effects of Relevance and Program were found; however, a significant interaction effect was found, $F(1, 54) = 4.445$, *MSE* = .819, *p* = .040, $\eta_p^2 = .076$. Visual inspection of this interaction effect in Figure 3.5 shows that, compared to the nursing group, only the care group takes advantage of being provided with the relevant criteria during the learning phase.

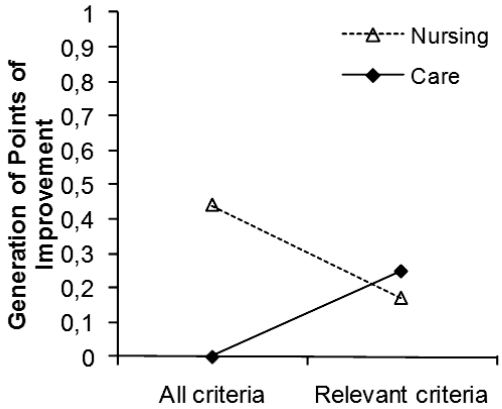

Figure 3.5: Interactions between Relevance and Program on generation of points of improvement.

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

Learning Phase

On average, students completed 3.36 learning tasks in the learning phase with a standard deviation of 3.64. There was thus a high variation with a minimum of 0 and a maximum of 14 learning tasks.

On learning task performance, a main effect of Program was found, $F(1, 54) = 42.416$, $MSE = 8.120$, $p = .00$, $\eta_p^2 = .44$. The care group ($M = 2.69$, $SD = .55$) showed a lower task performance than the nursing group ($M = 3.59$, $SD = .36$). Neither effect of Relevance, nor an interaction effect on learning task performance was found.

Neither significant effects of Relevance, Program, nor their interaction were found on number of tasks finished, accuracy of self-assessment, and reported mental effort for self-assessment.

Student Perceptions

Table 3.2 presents the means and standard deviations for the perception measures.

Table 3.2: 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

A main effect of Relevance on relevance of self-assessment was found, $F(1, 54) = 4.567$, $MSE = 1.874$, $p = .037$, $\eta_p^2 = .078$, indicating that students in the all criteria groups ($M = 3.04$, $SD = .60$) perceived the self-assessments as more relevant than students in the relevant criteria groups ($M = 2.68$, $SD = .67$).

On ability of self-assessment, a main effect of Program was found, F(1, 54) = 4.140, *MSE* = .723, *p* = .047, η_p^2 = .071, indicating that students in the care groups (*M* = 3.08, *SD* = .43) perceived they were better able to conduct self-assessments than students in the nursing groups (*M* = 2.76, *SD* = .41).

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

On task orientation a main effect of Program was found, 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 to have a clearer view on what was expected from them in the learning tasks than students in the nursing groups (*M* = 2.50, *SD* = .61).

On pleasure and interest in reflection, a main effect of Program was found, 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 to have more pleasure and interest in the reflection than students in the nursing groups (*M* = 1.63, *SD* = .60).

Finally, on usefulness, a main effect of Program was found, 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 the effects of drawing students' attention to relevant performance criteria on their task performance and the development of self-assessment skills. The first hypothesis, stating that students who receive the relevant criteria show higher test task performance than students who receive the undifferentiated list with criteria, was confirmed by our findings. The fact that students are better task performers when given the relevant criteria, validates the idea of Dochy et al. (1999) that providing them with the relevant criteria makes them understand better what good task performance looks like, which helps them in showing higher task performance. Furthermore, nursing students outperformed care students on task performance. This was a confirmation of the results in the learning phase, where nursing students also outperformed care students. A possible explanation is the different level of study program of nursing and care students. Nursing students are expected to have more insight in the processes of the nursing tasks and consequently to perform the tasks better.

The second hypothesis, stating that students who practice self-assessment skills with the relevant criteria will be better in selecting relevant criteria in the test phase than students who practice without the 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 provided in this study (van Merriënboer & Kir-

schner, 2007). The results did show that care students were able to select more relevant criteria than nursing students.

The third hypothesis, stating that students who receive the relevant criteria during the learning phase are better able to avoid the selection of irrelevant criteria in the test phase than students who merely receive the whole set of criteria, was not confirmed by the data. A possible explanation is that through highlighting, students' attention was only drawn to the relevant criteria. Students may have assessed their performance on the relevant criteria and simply ignored the irrelevant criteria during practice. It seems likely that paying more attention to the reasons behind the relevance of criteria during practice, for example, by explicitly asking students why some criteria are relevant and others are not, could have resulted in a better effect on the selection of relevant and irrelevant criteria. The results did 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 the relevance of the criteria.

The fourth hypothesis, stating that students who receive the relevant criteria will reach more accurate self-assessments than students who receive merely the whole set of criteria, was neither confirmed in the learning phase nor in the test phase. This is in contradiction with our expectations but in line with the findings of Dunning et al. (2004), who also found that for novice students knowing the relevant criteria does not necessarily imply the ability to self assess performance on those criteria. Possibly, more intensive practice would have been necessary.

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 the relevant criteria than nursing students; that is, highlighting increases the generation of points of improvement for the care group but not for the nursing group. However, this did not result in better test task performance for the care students. Apparently, they were not able to use their points of improvement to actually improve their performance.

With regard to cognitive load our results show no difference between groups during the learning phase, but students who received the relevant criteria during this phase reported a higher mental effort during the self-assessment in the test phase than students who received all criteria. Possibly, students who practiced with the relevant criteria are more engaged in finding the relevant criteria during the test phase, and therefore invest more effort in the self- assessment, in this case the load is positive for learning (Sweller et al., 1998). Another possible explanation is that students who practiced with the relevant criteria experience the assessment task in the test phase as more difficult than students in the all criteria group: For them, the test phase is the first time they are confronted with a whole undifferentiated list of

criteria. If this is the case, the load is negative for their learning (Sweller et al., 1998).

Next to these results, an intriguing finding is that students in the all criteria groups perceive the self-assessments as somewhat more relevant than students in the relevant criteria groups, possibly because they have to consider themselves which criteria are relevant and which are not. Furthermore, the care students were more positive in their perceptions. Concerning their own ability, they perceived their ability of self-assessment to be higher. Concerning the task, they indicated to have a clearer task orientation, they had more pleasure and interest in reflection, and they perceived the learning tasks as more useful. Overall, they seemed more positive than the nursing students but this was not reflected in their task performance.

Our study yields two practical implications. First, 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 each task as this helps students to perform the task better. Second, it seems fruitful to make a distinction between students who perform relatively low and students who perform relatively high on learning tasks (in our study, in order, care students and nursing students). For low-performing students, pointing out the relevant criteria seems especially important because it has positive effects on generating points of improvement, so that they know what to improve in future tasks.

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. For example, 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 beneficial for 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 irrelevant for assessing a particular task. In addition to drawing students' attention to relevant criteria, it would also be interesting to provide explanations why particular criteria are relevant and others are irrelevant, or to let students explain why particular criteria are relevant and others are not (i.e., 'self-explanation'; Renkl, 2002). Thinking-aloud protocols of students who verbally explain their thinking during self-assessment may also contribute to understanding the difficulties they encounter and the type of cognitive load they experience.

To conclude, the results of this study indicate that it is worthwhile to draw students' attention to the assessment criteria relevant for particular learning tasks. Highlighting these criteria enables students to perform better and to practice self-assessment skills with the appropriate criteria. Especially care students seem 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,* 121-132.

Boud, D., & Brew, A. (1995). Developing a typology for learner self assessment practices. *Research and Development in Higher Education, 18,* 130-135.

Boud, D., & Falchikov, N. (1989). Quantitative studies of student self-assessment in higher education: A critical analysis of findings. *Higher Education, 18,* 529-549.

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

Corbalan, G., Kester, L., & van Merriënboer, J. J. G. (2006). Towards a personalized task selection model with shared instructional control. *Instructional Science, 34,* 399-422.

Corbalan, G., Kester, L., & van Merriënboer, J. J. G. (2008). Selecting learning tasks: Effects of adaptation and shared control on efficiency and task involvement. *Contemporary Educational Psychology, 33*, 733-756.

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

Dochy, F., Segers, M., & Sluijsmans, D. (1999). The use of self-, peer and co-assessment in higher education: A review. *Studies in Higher Education, 24,* 331-350.

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*, 69-106.

Eva, K. W., & Regehr, G. (2005). Self-assessment in the health professions: A reformulation and research agenda. *Academic Medicine, 80*, 46-54.

Falchikov, N., & Boud, D. (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,* 386-391.

Gulikers, J. T. M., Bastiaens, T. J., Kirschner, P. A., & Kester., L. (2008). Authenticity is in the eye of the beholder: Student and teacher perceptions of assessment authenticity. *Journal of Vocational Education & Training, 60,* 401-412.

Jourdain, G., & Chênevert, D. (2010). Job demands-resources, burnout and intention to leave the nursing profession: A questionnaire survey. *International Journal of Nursing Studies, 47,* 709-722.

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.

Kicken, W., Brand-Gruwel, S., van Merriënboer, J. J. G., & Slot, W. (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*, 439-460.

Kittler, J. E., Menard, W., & Phillips, K. A. (2007). Weight concerns in individuals with body dysmorphic disorder. *Eating Behaviors, 8,* 115-120.

Kruger, J., & Dunning, D. (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,* 1121-1134.

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

Orsmond, P., Merry, S., & Reiling, K. (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,* 309-323.

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

Renkl, A. (2002). Worked-out examples: instructional explanations support learning by self-explanations. *Learning and Instruction, 12,* 529-556.

Rust, C., Price, M., & O'Donovan, B. (2003). Improving students' learning by developing their understanding of assessment criteria and processes. *Assessment & Evaluation in Higher Education, 28*(2), 147-164.

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

Sluijsmans, D. M. A., Straetmans, G. J. J. M., & van Merriënboer, J. J. G. (2008). Integrating authentic assessment with competence-based learning in vocational education: the Protocol Portfolio Scoring. *Journal of Vocational Education & Training, 60*, 159 -172.

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

Van Merriënboer, J. J. G. (1997). *Training complex cognitive skills.* Englewood Cliffs, NJ: Educational Technology Publications.

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.

Wolf, L. D., Potter, P., Sledge, J. A., Boxerman, S. B., Grayson, D., & Evanoff, B. (2006). Describing nurses' work: combining quantitative and qualitative analysis. *Human Factors, 48,* 5-14.

CHAPTER 4

The effects of performance-based assessment criteria on student performance and self-assessment skills

Abstract

This study investigated the effect of performance-based versus competence-based assessment criteria on task performance and self-assessment skills among 39 novice secondary vocational education students in the domain of Nursing and Care. In a performance-based assessment group students are provided with a preset list of performance-based assessment criteria, describing what students should do, for the task at hand. The performance-based group is compared to a competence-based assessment group in which students receive a preset list of competence-based assessment criteria, describing what students should be able to do. The test phase revealed that the performance-based group outperformed the competence-based group on test task performance. In addition, higher performance of the performance-based group was reached with lower reported mental effort during training, indicating a higher instructional efficiency for novice students.

This chapter is based on: Fastré, G. M. J., van der Klink, M. R., and 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,* 517-532.

In competence-based education, authentic learning tasks based on real-life problems are the driving force behind training, simultaneously encouraging the development of professional skills and more general competences like being self-directed. Competence-based education is a dominant trend in vocational education in many European countries (Wesselink, Biemans, Mulder, & van den Elsen, 2007). The aim is to prepare students for the workplace where people are expected to be broadly educated while stimulating lifelong learning (van Merriënboer, van der Klink, & Hendriks, 2002; van Merriënboer, Kirschner, Paas, Sloep, & Caniëls, 2009). Because competences are context-bound and the aim of vocational education is preparing students for the workplace, students should always develop competences in the context of a profession (Biemans, Nieuwenhuis, Poell, Mulder, & Wesselink, 2004). When teachers want to judge the competence development of their students, student assessments performed in a real-life context can support their findings.

Assessment criteria and standards are key clues for students to know what is essential in their study program. Fastré, van der Klink, Sluijsmans, and van Merriënboer (2010) show that drawing students' attention to the assessment criteria that are relevant for a particular learning task improves their understanding of the criteria and subsequently leads to better test task performance and better self-assessment skills. The following citation of Otter (1995) emphasizes the importance of being familiar with the relevant assessment criteria:

Describing and making clear and public what the learner is intended to achieve changes the nature of assessment from a tutor-led system with fuzzy objectives and undisclosed criteria, to a student-led system with greater emphasis on formative development and personal responsibility. (p. 45).

In the behavioural tradition of instruction and instructional design, assessment criteria were performance-based, meaning that they described the desired performance in terms of what the student has to do (e.g., Mager, 1984). With the introduction of competence-based education, assessment criteria are often formulated as competences, in terms of what the student is able to do. However, no research so far has investigated the effects of this introduction of competence-based assessment criteria. The main goal of this study is to investigate the effects of competence-based versus performance-based assessment criteria on learning, test task performance, and students' self-assessment skills.

The difference between performance-based and competence-based assessment criteria should be seen as a continuum, where on the one end assessment criteria are formulated as competences, which are an integration of knowledge, skills and attitudes; and on the other end assessment criteria are formulated as performance indicators. Performance-based criteria can be linked directly to competence-based criteria and vice versa as they complement each other. When discussing the continuum, the two extremes and their underlying

connection will be tackled. The discussion will be coupled to the level of experience students have as it can be assumed that students with different levels of experience will have different needs concerning assessment criteria (Kalyuga, 2007). In this chapter the focus is on the needs of novice students.

Figure 4.1 presents a summary of the continuum between competence-based and performance-based assessment criteria: (1) What is assessed, (2) the nature of the criteria, (3) holistic versus analytic, and (4) the level of mental effort.

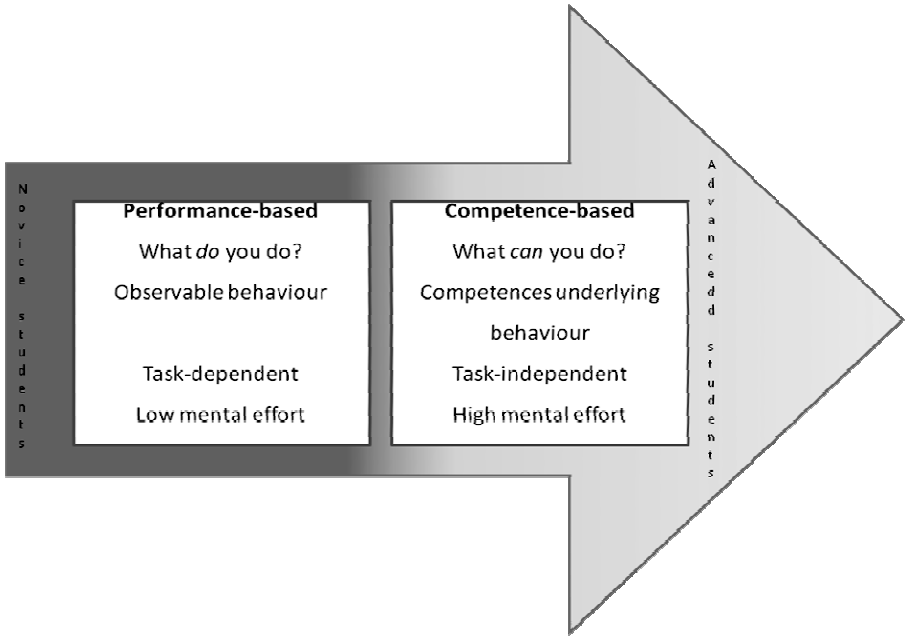

Fig 4.1: Continuum of performance-based to competence-based assessment criteria

First, with regard to what is assessed, when assessing with competence-based criteria, the competences underlying the performance are the focus of the assessment. What is assessed is the student's ability to perform a certain task. However, competences as a whole are not directly observable (Grégoire, 1997). Certain aspects of competences are observable, like particular skills the students demonstrate, but certain aspects are hidden, like their self-concept and personal characteristics that influence their performance (Spencer & Spencer, 1993).

When assessing with performance-based criteria, the observable behaviours produced by the students are the heart of the assessment. The question is not if the student is able to perform the task, but if the student shows good performance (Grégoire, 1997). In order to show this good performance, students probably also know how to perform and consequently master the underlying competences necessary for performing the task (Miller, 1990). For example, in the case of stoma

care, the student shows he can remove the stoma bag in a correct way. An underlying competence is supporting the patient according to protocols, regulations and the vision of the organisation but the performance criterion is removing the stoma bag in a correct way. This means there is a direct link between what students show (performance) and what students are able to do (competence). Every performance shown involves one or more competences the student has to possess to perform well, and every competence can be shown in several behaviours of the student.

Because for novice students it is important in an early stage to obtain an idea of how well they are doing, the directly observable character of the performance-based criteria may be expected to be more beneficial to assess their task performance. Based on these performance-based criteria, the development of the students from the beginning on can be monitored. In order to improve novice students' self-assessment skills, it is easier to assess what they are actually doing because this is more objective than their ability to do so. Therefore, with regard to what is assessed, performance-based criteria are expected to be more beneficial for supporting novice students' learning than competence-based criteria. In later stages, it is important for students to learn to see the link with the underlying abilities they are developing.

Second, with regard to the nature of the criteria, to uncover competence development, consistency of proof of competence level across different tasks is needed (Albanese, Mejicano, & Anderson, 2010; Grégoire, 1997). It is therefore important to formulate competence-based assessment criteria in a way that they can be used across different tasks and thus are task-independent. For example, a nurse has to be able to conduct nursing technical skills. In one situation this means replacing a stoma bag while in another situation this means washing a patient.

To judge student performance on a certain task, performance-based assessment criteria should be formulated on task-level as for each task a different set of criteria is relevant. Performance-based criteria are thus task-dependent. As is shown by Fastré et al. (2010), for novice students it is important to know the relevant criteria in every task. For example, when a nurse has to conduct stoma care, some of the relevant criteria are to remove the old stoma bag and apply a new one.

It is likely that when students know exactly what to do, their motivation, learning, and performance will increase significantly (see for example Ecclestone, 2001http://www.springerlink.com/content/3v4110m2n8u358p7/fulltext.html - CR5). Moreover, Miller (2003) argues that having task-specific assessment criteria leads to a better quantitative differentiation of performance levels. This more detailed view on students' performance, would argue for the use of performance-based assessment criteria. Following the results of Fastré et al. (2010), it can be

concluded that the use of performance-based criteria is especially beneficial for novice students because of their task-specific character.

Third, the competence-based assessment model currently used in Europe, starts from a fixed set of competences that are categorically divided (e.g., communication skills, nursing technical skills). No further decomposition of the competences is made. The formulation of the competence-based assessment criteria is therefore holistic (Grégoire, 1997). This does not mean that when working with competence-based assessment criteria only a holistic judgment on the end result is given, but the criteria are more holistically formulated than the performance-based criteria.

In a performance-based assessment model, the whole task is hierarchically analysed by developing a skills hierarchy (van Merriënboer, 1997; Sadler, 1985). In this way, criteria are expressed as a component of a higher-level criterion or a number of lower-level criteria. After the student performed the task, the teacher gives separate judgments on each of the preset criteria. Then, these judgments are combined to compose a final judgment which is often converted into a grade. As an example, Figure 4.2 shows a part of this decomposition for performing the task of stoma care.

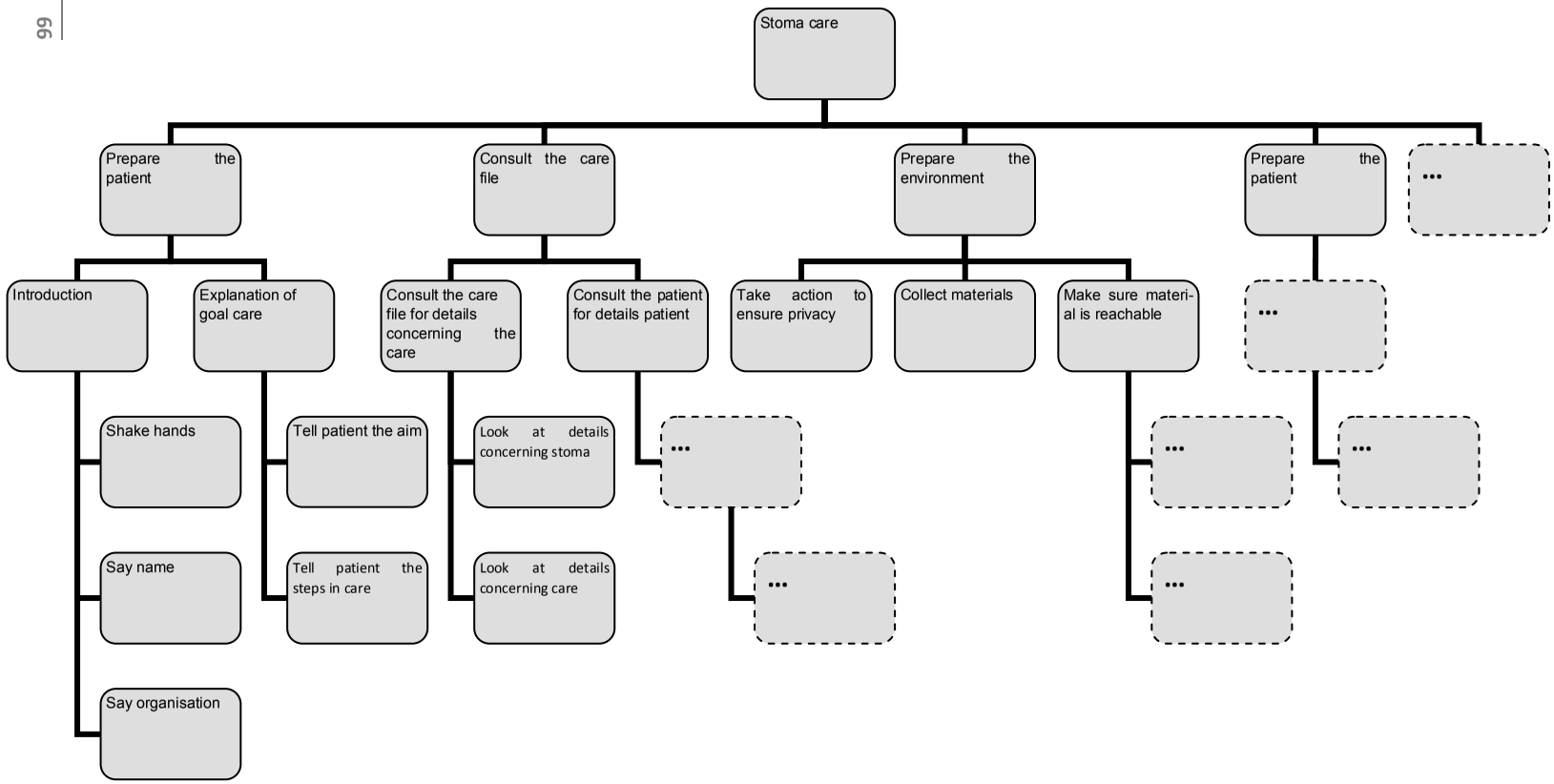

Figure 4.2: Part of the skill hierarchy of stoma care.

Gulikers, Kester, Kirschner, and Bastiaens (2008) discuss the notions of analytic versus holistic grading from the perspective of the level of experience of students. They argue that novice students need analytic criteria as guidelines in a step-by-step process leading to the desired behaviour. In future tasks, this helps to set appropriate learning goals (Eva & Regehr, 2005). For more advanced students, analytic criteria may hamper their learning process because they have to be stimulated to keep their focus on a certain outcome level and they do not need the step-by-step approach any more (Scheffer, Muehlinghaus, Froehmel, & Ortwein, 2008). Following these ideas, for novice students it would be better to receive performance-based assessment criteria.

Finally, with regard to mental effort, when designing a study program, including assessment, it is important to strive for the optimal level of using students' cognitive capacity (van Gog & Paas, 2008). Cognitive load theory presupposes that people have a limited working memory capacity (Sweller, van Merriënboer, & Paas, 1998; van Merriënboer & Sweller, 2005). Because of this limited capacity, it is essential for learning to properly allocate the available cognitive resources (Kalyuga, Ayres, Chandler, & Sweller, 2003).

An important difference can be distinguished here between novice students and more advanced students. For novice students, it is important to provide sufficient guidance that compensates for the limited knowledge they have on the task at hand (e.g., stoma care) by providing them performance-based assessment criteria because this requires less cognitive capacity for the assessment and most of their working memory capacity can be devoted to the task of stoma care. For more advanced students, who already have some knowledge on the task at hand (e.g., stoma care), competence-based assessment criteria can provide them with an extra stimulus to think about the task in another way and thereby make the extra cognitive capacity beneficial for them. In addition, providing these students with performance-based assessment criteria would give them redundant information on the task which may hamper their learning. This is called the expertise reversal effect (Kalyuga, 2007).

Summarising, it appears that for novice students, performance-based criteria have more advantages than competence-based criteria because: (1) They are directly observable, (2) they lead to a better quantitative differentiation of levels of performance, (3) they stimulate a step-by-step process leading to desired performance, and (4) they require less cognitive capacity for assessment leaving more capacity for learning the task at hand. The following section describes the hypotheses following this comparison.

Hypotheses

The first hypothesis is that students who receive the performance-based criteria during learning will show superior test task performance compared to students who receive the competence-based criteria because they know better what is expected from their performance. The second hypothesis is that students who receive the performance-based criteria will experience a lower mental effort during assessment than students who receive the competence-based criteria. The third hypothesis is that students who receive the performance-based criteria will be better self-assessors than students who receive the competence-based criteria because they are better able to assess their performance.

Method

Participants

Thirty-nine second-year students of a school for secondary vocational education, attending a Nursing and Care program (Level 3 and 4 in the European Qualifications Framework, 2 males and 37 females) participated in this study as part of their regular training on the nursing task of stoma care. Their mean age was 18.07 years (*SD* = 1.05). Participants were randomly assigned to one of the two conditions: competence-based criteria (*n* = 20) and performance-based criteria (*n* = 19).

Materials

The whole task of stoma care, addressing the psychosocial needs of the patient, analysing the situation of the patient, changing the stoma bag, and the evaluation afterwards are included in the task. This means students did not only practice the technical skill of changing the stoma bag, but also needed knowledge on the stoma (e.g., possible problems with stomas), and an appropriate attitude towards the patient. The task was set up in accordance with the theory of experiential learning by Steinaker and Bell (1979) which distinguishes four important steps: (a) exposure, (b) participation, (c) identification, and (d) internalisation. Figure 4.3 summarises the materials described below.

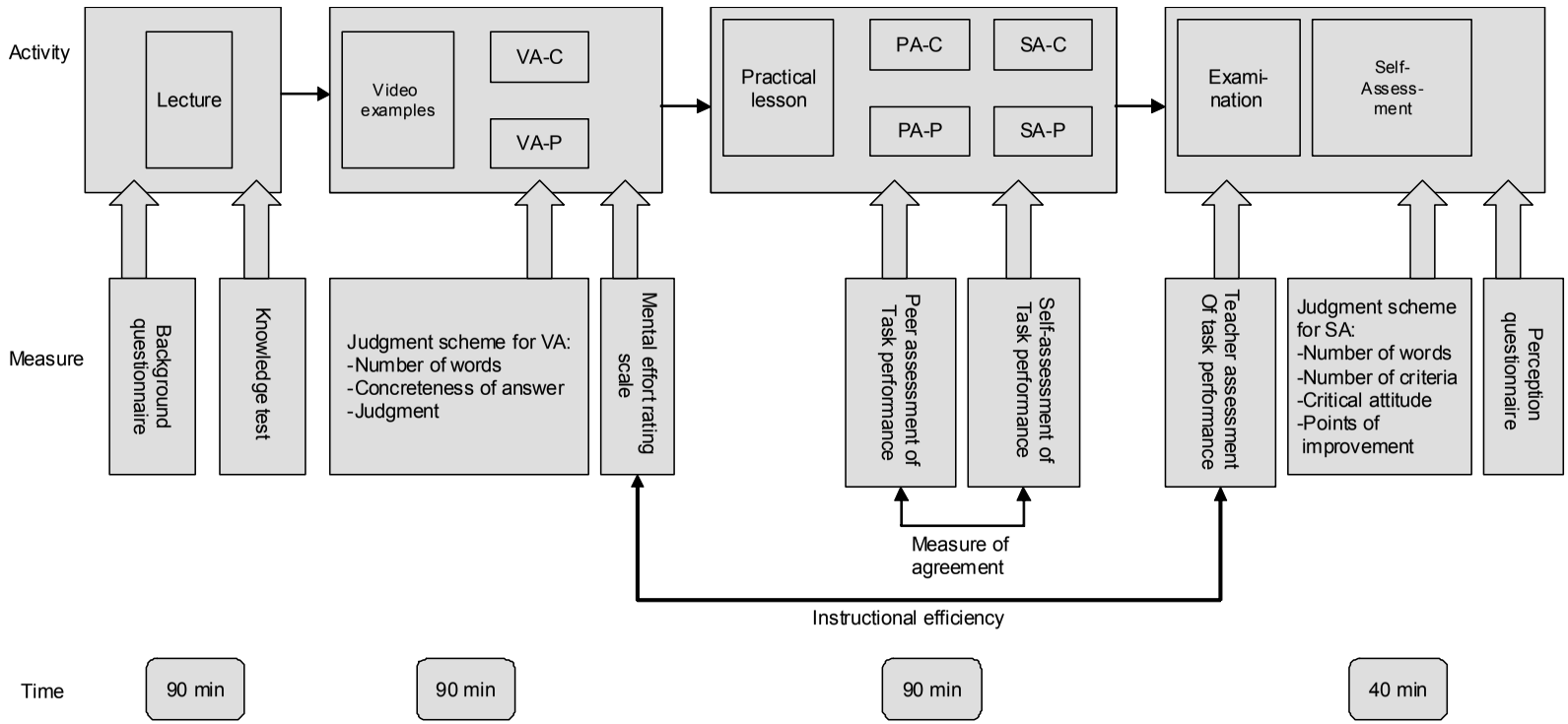

Figure 4.3: Summary of materials and measures.

Lecture

A lecture was developed that provided students with the theoretical background of stoma care. The two teachers who were responsible for this lecture set up the lecture together.

Video examples and video assessment

An electronic learning environment was developed including six video fragments (±3 min each) in which an expert nurse shows good stoma care behavior. All fragments are subsequent parts of the whole task of stoma care: (1) introduction, (2) preparation, (3) removing the old stoma bag, (4) applying the new stoma bag, (5) finishing off care, (6) evaluation and reporting. Students individually watched the video fragments on a computer screen. They were not allowed to put the fragment on hold, and they could watch the video a maximum of three times. On average, students watched the video 1.14 times (*SD* = .29). No differences between conditions were found.

After students watched the video, they had to assess the performance of the nurse in the video on an electronic list of preset criteria. A distinctive feature was made for the two conditions. In the competence-based condition, the assessment criteria were formulated as competences of stoma care as used previously in the study program (VA-C). Figure 4.4 shows some examples of competence-based criteria as shown in the electronic learning environment.

Criterion		How does the nurse show this criterion
Giving attention and sympathising	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 pro-active	
	Consults patients and involved others on a regular basis and informs them	
	Makes agreements with patients and others involved in the task division	
Ethical and honest treatment	Acts consequent to own norms, professional group, organisation and legal constraints	
	Is integer, fair and acts without prejudice	
	Is discrete with sensitive topics	
	Communicates open and clear about sensitive topics	

Figure 4.4: Screen dump of competence-based assessment criteria.

In the performance-based condition, the assessment criteria were formulated as the underlying skills of a skill hierarchy of stoma care (VA-P). Figure 4.5 shows some examples of performance-based criteria as shown in the learning environment.

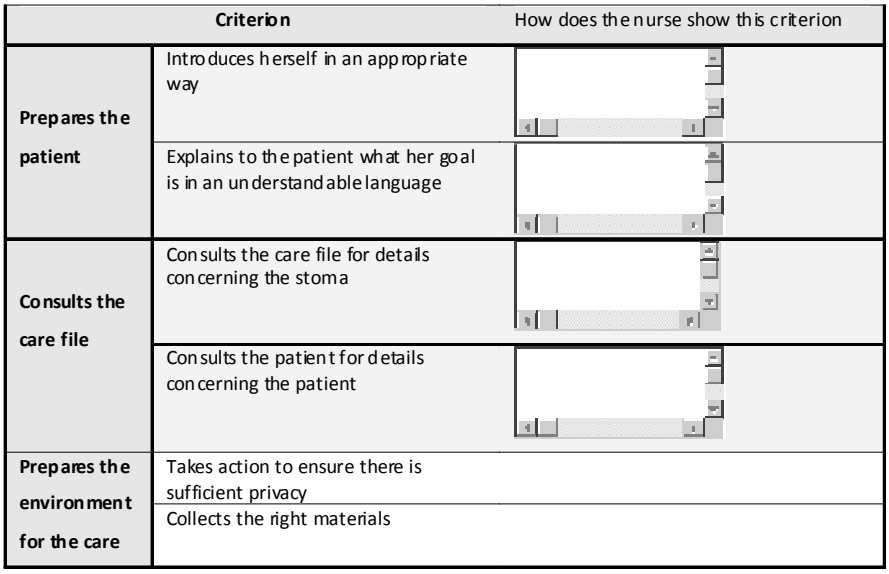

Figure 4.5: Screen dump of performance-based assessment criteria.

In order to encourage students to make the assessment criteria more concrete, students in both groups had to indicate the manner in which the nurse in the fragment showed good behaviour on the criteria by typing their answer in the text boxes.

Practical lesson, peer assessment, and self-assessment

A practical training session was developed in which students had to practice in pairs or groups of three the task of stoma care with a fellow student being the patient. After students had performed the task, they had to score their peers' task performance on the same list of criteria as in the assessment of the video examples. The students in the competence-based condition received the list with competence-based criteria (PA-C) and students in the performance-based condition received the list with performance-based criteria (PA-P). They had to indicate how well their peers mastered the criteria on a four-point scale: (1) behaviour not shown, (2) behaviour shown but insufficient, (3) behaviour shown and sufficient, (4) behaviour shown and good. In addition to this peer assessment, students had to self asses their task performance using the identical list of competence-based criteria (SA-C) or performance based criteria (SA-P), using the same four-point scale. While

practicing the task, students also received oral feedback on their task performance from the instructor in the room.

Examination and self-assessment
An examination was developed in which students individually had to perform the task of stoma care with a simulation patient. Afterwards they had to assess their own performance on that particular task by filling in a blank paper with the question: assess your own performance on this task and indicate what went well and what went wrong.

Measures

Background questionnaire
A short questionnaire measured the background of the students on demographical factors such as age, sex, and prior education. Student perceptions of the relevance of self-assessment and their perceptions of their ability to self assess were measured by the self-directed learning skills questionnaire adapted from Kicken, Brand-Gruwel, and van Merriënboer (2006). This questionnaire proved reliable for the population in this study (Fastré et al., 2010). Table 4.1 shows the Cronbach's alpha scores of the perception scales; internal consistencies ranged from .70 to .75 and are thus quite acceptable.

Table 4.1: Reliability of the Self-Directed Learning Skills Questionnaire.

Scale	Cronbach's Alpha	# items	Example item[a]
Self-Directed Learning Skills Questionnaire			
Relevance of self-assessment	.70	3	I think the opinion of the teacher is more important than self-assessment in Care Village.
Ability to self assess	.75	12	I can assess to what extent my performance fits the assessment criteria in Care Village.

[a] Items have been translated from Dutch.

Knowledge test
At the end of the lecture, a 15-item multiple choice test was taken to test the students' knowledge on this subject.

Judgment scheme for video assessment
To measure the accuracy of the video assessment, judgment schemes specified the quality of the video assessments. The overall score for quality of video assessment was the sum of the *z*-scores of the following aspects: how many words the students used because it is expected that performance-based criteria stimulate students more to elaborate on their answers (count of the number of words), if they gave

concrete examples of the nurse's behaviour (0 = no concrete behaviour, 1 = concrete behaviour), and if they gave a judgment on the behaviour of the nurse (0 = no judgment, 1 = judgment). The higher the sum of the *z*-scores, the better the score for quality of video assessment as it is important that the combination of these factors is of a high quality. The quality of the video assessments was judged by two raters, with a high interrater reliability of $r = .82$, $p < .00$.

Mental effort rating scale

After the assessment of each video fragment, 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, Kester, and van Merriënboer (2009). The scale points ranged from very low mental effort (1) to very high mental effort (7).

Peer assessment of task performance

The peer assessments during the practical lesson indicated the task performance of the students assessed by the peers, using the competence-based criteria in one group and performance-based criteria in the other group. Peer assessed task performance was the average score on all the assessment criteria.

Self-assessment of task performance

The self-assessments during the practical lesson indicated the task performance of the students by the students' own opinion, using the competence-based criteria in one group and performance based criteria in the other group. Self assessed task performance was the average score on all the assessment criteria.

Teacher assessment of test task performance

During the examination, the teachers observed and assessed the test task performance of the students, who took care of the stoma of a simulation patient, on the list of performance-based criteria. A second assessor co-assessed with each of the teachers to measure the reliability of the assessments. The correlation between the scores of the teacher and the second assessor, $r = .77$, $p < .01$, appeared to be acceptable.

Judgment scheme for self-assessment

The overall score for quality of the self-assessments during examination was the sum of the *z*-scores of the following aspects: how many words the students used because it is expected that performance-based criteria stimulate students more to elaborate when self assessing (count of the number of words), how many criteria the students came up with (count of the number of criteria), if students had a critical attitude to their own performance (0 = no critical attitude, 1 = critical attitude), and if they formulated points of improvement (0 = no points of

improvement, 1 = points of improvement). The higher the sum of the *z*-scores, the better the score for quality of self-assessment because it is important that the combination of these factors is of a high quality. The quality of the self-assessments was judged by two raters, with an interrater reliability of $r = .82$, $p < .00$.

Perception questionnaire

The following aspects of perception were measured to evaluate the learning experience: Motivation for the study, regulation strategies, interesting course material, task orientation, pleasure and interest, pleasure and interest in reflection, and usefulness. All aspects were measured with the use of four-point Likert scales. Higher scores indicate a more positive perception of the learning experience. Two scales (interesting course material and task orientation) of the inventory of perceived study environment (IPSE; Wierstra, Kanselaar, van der Linden, & Lodewijks, 1999) measured students' perceptions of the learning environment. Three scales (interest and pleasure, interest and pleasure in reflection, and usefulness) of the intrinsic motivation inventory by Deci, Eghrari, Patrick, and Leone (1994), translated into Dutch by Martens and Kirschner (2004), were included in the questionnaire. Table 4.2 shows the Cronbach's alpha scores of the perception scales; internal consistencies ranged from .69 to .89 and are thus acceptable to high.

Table 4.2: Reliability of the perception measures.

Scale	Cronbach's Alpha	# items	Example item[a]
Inventory of Perceived Study Environment			
Interesting course materials	.72	8	The learning tasks are interesting.
Task orientation	.69	3	I know what is expected of me when performing a task.
Intrinsic Motivation Inventory			
Interest and pleasure in learning tasks	.69	7	I enjoy working on the learning tasks.
Interest and pleasure in reflection	.89	6	I find it interesting to reflect.
Usefulness	.69	4	I would like to conduct more learning tasks because they are useful.

[a] Items have been translated from Dutch.

Measure of agreement

For the peer assessments and the self-assessments during the practical lesson, the agreement of the scores between the self- and peer assessments was measured by computing the Pearson's correlation.

Instructional efficiency

Instructional efficiency is calculated by relating task performance in the test task and experienced mental effort during training (Paas & van Merriënboer, 1993; van

Gog & Paas, 2008). Performance and mental effort scores are first standardized, and then the *z*-scores are entered into the formula:

$$E = \frac{Z_{Performance} - Z_{MentalEffort}}{\sqrt{2}}$$

High efficiency indicates that with a relatively low mental effort during training a relatively high task performance in the examination is accomplished, while a low efficiency indicates that with a relatively high mental effort during training a relatively low task performance is accomplished. For example, instructional efficiency is higher for an instructional condition in which participants attain a certain performance level with a minimum investment of mental effort than for an instructional condition in which participants attain the same level of performance with a maximum investment of mental effort.

Procedure

At the start of the lecture, the background questionnaire was administered. After students had filled in the questionnaire, the lecture was given and the multiple choice test was taken. This phase lasted for 90 min.

After the lecture students were instructed to assess the video examples. While doing this, students were exposed to the stoma care by watching video examples of an expert nurse showing the intended behaviour, which is the first step in the taxonomy of Steinaker and Bell (1979). Students were split up in the two experimental groups to work on the assessment of video examples. Students could work on the assessment of video examples for maximum 90 minutes. After the assessment of video examples, the practical lesson with peer and self-assessments took place for 90 minutes. In this lesson, students could participate in stoma care by practicing on a doll (second step).

One week after the practical lesson, students had to conduct the examination after which they had to assess their own performance. In this examination, they could identify with the stoma care because they were exposed to a simulation patient in performing the care (third step). Student performance was assessed by a teacher. At the end of the examination the evaluation questionnaire was filled in by the students. The examination including self-assessment lasted for 40 minutes. After the whole experiment, students were sufficiently prepared for further practice during internships which leads them to internalise the competence of stoma care (fourth step).

Results

This section describes the results on prior measurements, the dependent variables in the learning and test phase, and the student perceptions. Mann-whitney *U* tests were performed to test for differences between the two conditions. For all analyses, the significance level is set to .05. Table 4.3 presents the means and standard deviations for all dependent variables.

Table 4.3: Means and standard deviations for dependent variables.

	Competence-based (*n* = 20)		**Performance-based (*n* = 19)**	
	M	*SD*	*M*	*SD*
	Pretest			
Score on prior knowledge test	7.72	1.27	7.85	1.35
	Learning phase			
Score quality of video assessment	.08	1.50	.22	1.65
Number of words	889	143	855	248
Concreteness of Answer	.50	.52	.80	.41
Judgment	.56	.51	.43	.51
Mental effort	3.33	.75	2.36	.65
Task performance scored by peer	2.95	.34	3.14	.43
Task performance scored by self	3.12	.31	3.25	.48
	Test phase			
Test task performance scored by teacher	2.89	.44	3.18	.45
Score quality of self-assessment	-.53	2.73	.56	2.92
Number of words	93	45	111	67
Number of criteria	6.35	3.80	6.32	3.86
Critical attitude	.60	.50	.74	.45
Points of improvement	.15	.37	.37	.50
Instructional efficiency	-.65	.81	.68	.91

Prior measurements

On the background questionnaire, no significant difference between the conditions was found, indicating that students did not differ in background at the end of the lecture.

On the knowledge test, no significant difference between the conditions was found, indicating that all students had the same level of knowledge at the end of the

lecture. Thus, students had the same background and prior knowledge before they started to study the video examples.

Learning phase

On the overall score for quality of video assessment, a significant difference between the conditions was found, $z = -1.964$, $p < .05$. Students in the performance-based condition had an average rank of 18.21, while students in the competence-based condition had an average rank of 12.00. More specifically, on number of words no difference was found. In concreteness of answers, a significant difference was found, $z = -1.716$, $p < .05$. Students in the performance-based condition had an average rank of 18.40, while students in the competence-based condition had an average rank of 13.75. No significant difference in judgment was found. A further qualitative analysis of the data reveals that students in the competence-based condition often decoded the competence-based assessment criteria into the performance-based criteria as an answer but were not able to describe the concrete behaviour.

Mental effort during assessment of the video examples is an average score of the scores during assessment of the six film fragments. On mental effort, a significant difference between conditions was found, $z = -3.964$, $p < .001$, indicating that students in the performance-based condition had an average rank of 12.61, while students in the competence-based condition had an average rank of 27.03.

On peer assessment and self-assessment of task performance in the practical lesson, no significant differences between conditions was found. Yet, a moderate agreement between peer and self-assessment was found, $r = .65$, $p < .00$, indicating that students' self-assessment scores corresponded with the scores of their peers. For the performance-based condition $r = .66$, $p < .01$, and for the competence-based condition $r = .63$, $p < .01$.

Test phase

On test task performance, a significant difference between conditions was found, $z = -2.037$, $p < .05$. Students in the performance-based condition had an average rank of 23.82, while students in the competence-based condition had an average rank of 16.38. On the overall score for quality of self-assessment, no significant differences between both conditions were found. Although not significant, the direction of the differences was in line with the expectations. On instructional efficiency, a significant difference between conditions was found, $z = -3.962$, $p < .001$, indicating that students in the performance-based condition had an average rank of 27.42, while students in the competence-based condition had an average rank of 12.95.

Evaluation questionnaire

Overall, students perceived the learning environment as interesting and useful. Table 4.4 shows the means and standard deviations for all scales.

Table 4.4: Means and standard deviations for evaluation questionnaire.

	Competence-based (*n* = 20)		**Performance-based (*n* = 19)**	
	M	*SD*	*M*	*SD*
Interesting course material	3.13	.43	3.07	.39
Task orientation	2.83	.67	2.51	.70
Interest and pleasure	3.47	.39	3.35	.37
Interest and pleasure in reflection	2.59	.43	2.51	.58
Usefulness	3.45	.52	3.50	.33

No significant differences were found between conditions. Being in the performance-based or competence-based condition did not influence students' perceptions of the learning task.

Discussion

The goal of this study was to investigate the effects of competence-based versus performance-based assessment criteria on students' test task performance and self-assessment skills. The first hypothesis, stating that students who receive the performance-based criteria will be better task performers than students who receive the competence-based criteria is confirmed by the data. It seems that novice students who receive the performance-based criteria during training know better what is expected from their task performance and are better able to show desired performance than students who receive the competence-based criteria. A possible explanation is the finding that students who receive the performance-based criteria had a higher quality of video assessments in the learning phase. They were especially better in being concrete on the desired behaviour, which may have led to better task performance in the test phase. This is in line with the ideas of Eva and Regehr (2005*),* who state that performance-based criteria make it easier to distinguish levels of performance, enabling a step-by-step process of performance improvement.

The second hypothesis, stating that students who receive the performance-based criteria experience a lower mental effort during assessment than students who receive the competence-based criteria is also confirmed by the data. It appears that by providing novice students with performance-based assessment criteria, they have to invest less mental effort to assess their task performance. This effect is positive when it leads to a better test task performance because this would mean

that during training the reduced load of assessment permits more cognitive capacity for learning to perform the task of stoma care.

Indeed, the findings concerning the first and second hypotheses together allow to conclude that the performance-based assessment criteria result into a higher instructional efficiency, since students in the performance-based condition experience a lower cognitive load during the learning phase, followed by a higher performance on the test task (Paas & van Merriënboer, 1993; van Gog & Paas, 2008). Providing novice students with performance-based assessment criteria thus leads to more efficient learning.

The third hypothesis, stating that students who receive the performance-based criteria become better self-assessors than students who receive the competence-based criteria, is not confirmed by the results. This finding is, however, in line with the findings of Dunning, Heath, and Suls (2004), who also found that for novice students knowing the assessment criteria does not necessarily imply the ability to assess their own performance on those criteria. As self-assessment can be seen as a complex cognitive skill, one of the key words in developing this skill is sufficient practice (van Merriënboer & Kirschner 2007). It is likely that students need considerably more practice than provided in the current study to improve their self-assessment skills.

Finally, students did not differ in their perceptions of the learning environment. Receiving competence-based or performance-based criteria thus did not influence their appreciation of the learning task. The findings indicate that both groups were positive about the learning task as a whole and especially valued the provided video examples.

The results of this study show that for novice students performance-based assessment criteria do lead to a lower mental effort during learning and a higher test task performance, which is in line with our theoretical assumption that for novice learners it is better to use performance-based criteria than competence-based criteria. The question remains, however, what causes the observed effects. The relative importance of the separate dimensions of Fig. 4.1 was not investigated in this study and further research is required to determine the contribution of the various dimensions to the reported effects on mental effort during learning and test task performance. Is it because these criteria refer to directly observable behaviour? Or is it because the criteria are more task-dependent? Maybe the analytic character of the criteria is the driving force behind these effects? These insights could serve as a guideline for teachers in the development of performance-based assessment criteria and should be further examined.

Furthermore, the effects of providing students with performance-based assessment criteria should be examined with students in later years of the educational program to explore differences between novice and more advanced students as it is expected that students in later phases of their educational program

have to learn to think on a higher level and thus work more efficient with competence-based criteria.

A shortcoming of this study is the limited duration of the intervention. Because this intervention was restricted to only one learning task (i.e., stoma care), students did not get the opportunity to practice extensively on their skill development. This was most visible for the complex cognitive skill of self-assessment. According to van Merriënboer and Kirschner (2007), more training is needed to develop this kind of skill. Furthermore, only a small sample was used in the study. The question remains if the results are transferable to larger groups of students or students in other domains. Nevertheless, the fact that this intervention yielded some important results concerning mental effort expenditure during learning and test task performance is a sound basis for further research on this topic.

The findings yield the clear guideline that novice students should be provided with performance-based assessment criteria in order to improve their learning process, and reach higher test task performance. For instructing young nurses in the beginning of their study, performance-based assessment criteria are a necessity to guide their learning process. It should be noted, however, that formulating such performance-based criteria is a demanding task. To assure a sound implementation, training should be provided to teachers to increase their skills in formulating performance-based assessment criteria, based on a systematic process of drawing up a skills hierarchy with related criteria. When students progress in the study program, explicit attention should be paid to training students to interpret their own behaviours in terms of the underlying competences. In this way, students learn to see the link between performance and competence development. If this is not explicitly in the program, students remain on a lower level of thinking.

To conclude, the introduction of competence-based education primarily consisting of authentic learning tasks based on real-life problems, leads educators to solve the issue of how to redesign their assessment programs. Our results show that stating that competence-based assessment criteria are the answer to this problem is a step too far. Whereas competences seem to be a good starting point to develop professional education, they do not always serve this purpose for assessment. At least for novice students, providing them with performance-based assessment criteria is more beneficial than providing them with competence-based criteria. This study shows that novice students need less mental effort to assess their task performance and show higher test task performance, that is, they learn more efficiently when being provided with performance-based assessment criteria.

References

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

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, 36*, 523-538.

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.

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

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*, 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*, 301-313.

Eva, K. W., & Regehr, G. (2005). Self-assessment in the health professions: A reformulation and research agenda. *Academic Medicine, 80*, 46-54.

Fastré, G. M. J., van der Klink, M. R., Sluijsmans, D., & van Merriënboer, J. J. G. (2010). *Drawing students' attention to relevant assessment criteria: Effects on self-assessment skills and performance.* Manuscript submitted for publication.

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

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*, 172-186.

Kalyuga, S. (2007). Expertise reversal effect and its implications for learner-tailored instruction. *Educational Psychology Review, 19*, 509–539.

Kalyuga, S., Ayres, P., Chandler, P., & Sweller, J. (2003). The expertise reversal effect. *Educational Psychologist, 38*, 23–31.

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

Mager, R. F. (1984). *Preparing instructional objectives* (2nd ed.). Belmont, CA: David S. Lake.

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*(Suppl.), S63-S67.

Miller, P. J. (2003). The effect of scoring criteria specificity on peer and self-assessment. *Assessment and Evaluation in Higher Education, 28*, 383-394.

Otter, S. (1995). Assessing competence: The experience of the enterprise in higher education initiative. In: Edwards, A. & Knight, P. (Eds.) *Assessing competence in higher education.* London: Kogan Page.

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

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

Scheffer, S., Muehlinghaus, I., Froehmel, A., & Ortwein, H. (2008). Assessing students' communication skills: Validation of a global rating. *Advances in Health Sciences Education, 13,* 583-592.

Spencer, L., & Spencer, S. (1993). *Competence at work*. New York: John Wiley&Sons, Inc.

Steinaker, N. W., & Bell, M. R. (1979). *The experiential taxonomy: A new approach to teaching and learning*. New York: Academic Press.

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

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. (1997). *Training complex cognitive skills*. Englewood Cliffs, NJ: Educational Technology Publications.

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., Kirschner, P. A., Paas, F., Sloep, P. B., & Caniëls, M. C. J. (2009). Towards an integrated approach for research on lifelong learning. *Educational Technology, 49*, 3-14.

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.

Van Merriënboer, J. J. G., van der Klink, M. R., & Hendriks, M. (2002). Competenties: Van complicaties tot compromis [Competencies: From complications to compromise]. The Hague: Educational Council.

Wesselink, R., Biemans, H. J. A., Mulder, M., & van den Elsen, E. R. (2007). Competence-based VET as seen by Dutch researchers. *European Journal of Vocational Training, 40*, 38–50.

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.

CHAPTER 5

What are effective assessment criteria for competence-based education?

Abstract

This study examined the effects of type of assessment criteria (performance-based vs. competence-based), relevance of assessment criteria (relevant criteria vs. all criteria), and their interaction on secondary vocational education students' performance and assessment skills. Students of three programs in the domain of Nursing and Care (*N* = 93) participated in the study. Results show that students who received the relevant criteria made more accurate assessments of an expert model, performed better on a test, and achieved higher instructional efficiency compared to students who received all criteria. Students who received performance-based assessment criteria made more accurate assessments of an expert model and scored higher on task performance during practice. They invested less mental effort in the assessments, resulting in higher instructional efficiency. An interaction effect for the concreteness of assessments shows that the combination of performance-based and relevant criteria leads to superior performance compared to the other combinations of criteria.

This chapter is based on: Fastré, G. M. J., van der Klink, M. R., Amsing-Smit, P., and van Merriënboer, J. J. G. (2010). *What are effective assessment criteria for competence-based education?* Manuscript submitted for publication.

Both nursing students and car mechanic students have to be able to analyze problems, cooperate with others and present information. However, the contexts in which they have to demonstrate these competences differ significantly. A student nurse has to be able to analyze a patient file while a car mechanic student has to be able to analyze a car defect. Nevertheless, with the recent introduction of competence-based education in secondary vocational education in the Netherlands, profiles have been composed for core tasks, work processes and competences that are identical for all programs, even though these programs are aimed at training competent professionals who are well equipped to deal with future developments in very different professional fields (Biemans, Nieuwenhuis, Poell, Mulder, & Wesselink, 2004). Although competences differ as a function of work context, a fixed set of 25 general competences has been developed which students must have mastered at the end of all educational programs in secondary vocational education (COLO, 2008).

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

For novice students in particular, assessment criteria are important clues to determine the core content of their study program (Sadler, 1985). Students should therefore be provided transparent assessment criteria before they tackle learning tasks. In many Dutch secondary vocational education programs, students are required to select competences they want to master from the above-mentioned list of 25 general competences, which are also used as assessment criteria. It is questionable, however, if these general, competence-based assessment criteria offer novice students a sufficiently sound basis for assessing their own performance on learning tasks, since these competences are not tailored to the work context of their future profession. These questions regarding competence-based assessment criteria coupled with the fact that the nature of effective assessment criteria in competence-based education is an under-researched topic (Fastré, van der Klink, & van Merriënboer, 2010) led to the study reported in this chapter. The main research question is: which assessment criteria are most effective in promoting student learning.

The study focused on two major differences between criteria: competence-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.

Competence-based and Performance-based Criteria

Within competence-based education, assessment criteria are often formulated as competences, in other words, what the student is *able* to do (Grégoire, 1997). 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*. The difference between these two types of criteria should be seen as a continuum which is linked to different points of a training program, in other words they are related to different levels of training. Competence-based criteria are less meaningful for novice students than for advanced students, because novices have not yet achieved the requisite integration of knowledge, skills, and attitudes (Fastré, van der Klink, & van Merriënboer, 2010). There is, however, a direct relationship between competence- and performance-based criteria, because the latter specify context-specific performance in relation to the competence in question. For student nurses, for example, the performance-based criteria related to the competence 'communicating properly' are 'introduce yourself to the patient', 'tell the patient what you are going to do', et cetera.

Figure 5.1 shows the continuum from performance-based to competence-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' and the three others are: (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.

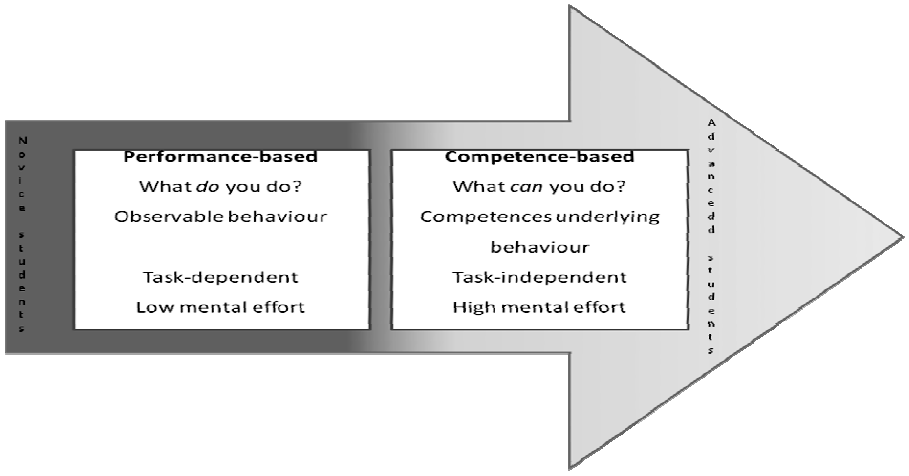

Figure 5.1: Continuum of performance-based to competence-based assessment criteria.

Competence-based criteria address the student's ability to perform a certain task rather than actual task performance. As a result students first have to *interpret* the criteria to relate them to performance before they can make an accurate assessment (Grégoire, 1997). 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 competences (Miller, 1990). Performance-based criteria are expected to be more beneficial for novice students than competence-based criteria, because they do not require the prior knowledge that is indispensable for the correct interpretation of competence-based criteria.

The knowledge that is needed to link competence with performance can only be developed from experience with a series of different tasks. Furthermore, competences go beyond concrete task performance and are best demonstrated through acceptable performance of a variety of tasks. Thus, competence-based criteria are not dependent on a particular task (Albanese, Mejicano, & Anderson, 2010), unlike performance-based criteria, which are task-dependent. A different set of performance-based criteria may be relevant for each new task. Performance-based criteria are expected to be more beneficial for 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).

Interpretation of competence-based criteria requires mental effort, thereby increasing the cognitive load for 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, van Merriënboer, & Paas, 1998; van Merriënboer & Sweller, 2005). Performance-based criteria are less costly in terms of cognitive load, because they do not require the cumbersome process of interpretation before the actual assessment. This is particularly relevant for novice students, who have to expend considerable mental effort to link competence and performance because they have not yet acquired the necessary knowledge to translate competence to performance. Performance-based criteria reduce cognitive load, because they are directly observable and specify concretely what is expected from students. Earlier research by Fastré, van der Klink, and van Merriënboer (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 received competence-based assessment criteria.

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

Relevance of Criteria

For complex learning tasks, 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 split into two parts for each learning task: Relevant and irrelevant criteria. Students in competence-based vocational education are often given long lists of pre-specified criteria, such as the 25 general competences mentioned earlier, from which they have to select the criteria that are relevant to the task at hand (Kicken, Brand-Gruwel, & van Merriënboer, 2008). This practice is based on the notion that for students to become independent learners they must learn to make the distinction between relevant and irrelevant criteria for the tasks they undertake.

However, being confronted with a long list of all potentially relevant criteria may have negative consequences for novice students, who are insufficiently equipped to identify which criteria are and which are not relevant for a specific task (Regehr & Eva, 2006; Dunning, Heath, & Suls, 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 may 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, van der Klink, Sluijsmans, and van Merriënboer (2010), in which novice students who received 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 expected that presenting students with relevant performance-based assessment criteria would be most beneficial for their learning. Additionally, it is possible that specific combinations of different criteria can contribute to significant learning gains. It seems a reasonable assumption that students who are given a long list of all potentially relevant assessment criteria stand to gain more from performance-based criteria, which are more concrete than competence-based criteria and whose relevance is therefore easier to determine. It also seems reasonable to assume that students who receive competence-based criteria profit more from relevant criteria than from a list of all possible criteria, because relevant criteria make it easier to link competences to specific performance. In order to examine these assumptions, we explored whether some combinations of criteria of different types

and relevance induced better task performance, superior assessment skills, and lower investment of mental effort compared to other combinations.

It is hypothesized that students who receive performance-based criteria will show better task performance, better assessment skills, and lower investment of mental effort than students who receive competence-based criteria. Second, it is hypothesized that students who receive only relevant assessment criteria will show better task performance and better assessment skills than students who receive all potentially relevant assessment criteria. Additionally, interaction effects of relevance and type of criteria are explored to test the hypothesis that a combination of relevant and performance-based criteria is most conducive to learning.

Method

Participants

Ninety-three second-year students attending Nursing and Care programs at three different institutes of secondary vocational education (7 males and 86 females) participated in this study as part of their regular training in stoma care. The mean age was 18.03 years (*SD* = 1.01) and the number of participants per institution was 32, 39, and 22 respectively. In each institute, participants were randomly assigned to one of four conditions: Competence-based/all criteria (*n* = 23), competence-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 in all institutes.

Materials

Figure 5.2 summarizes the materials described below.

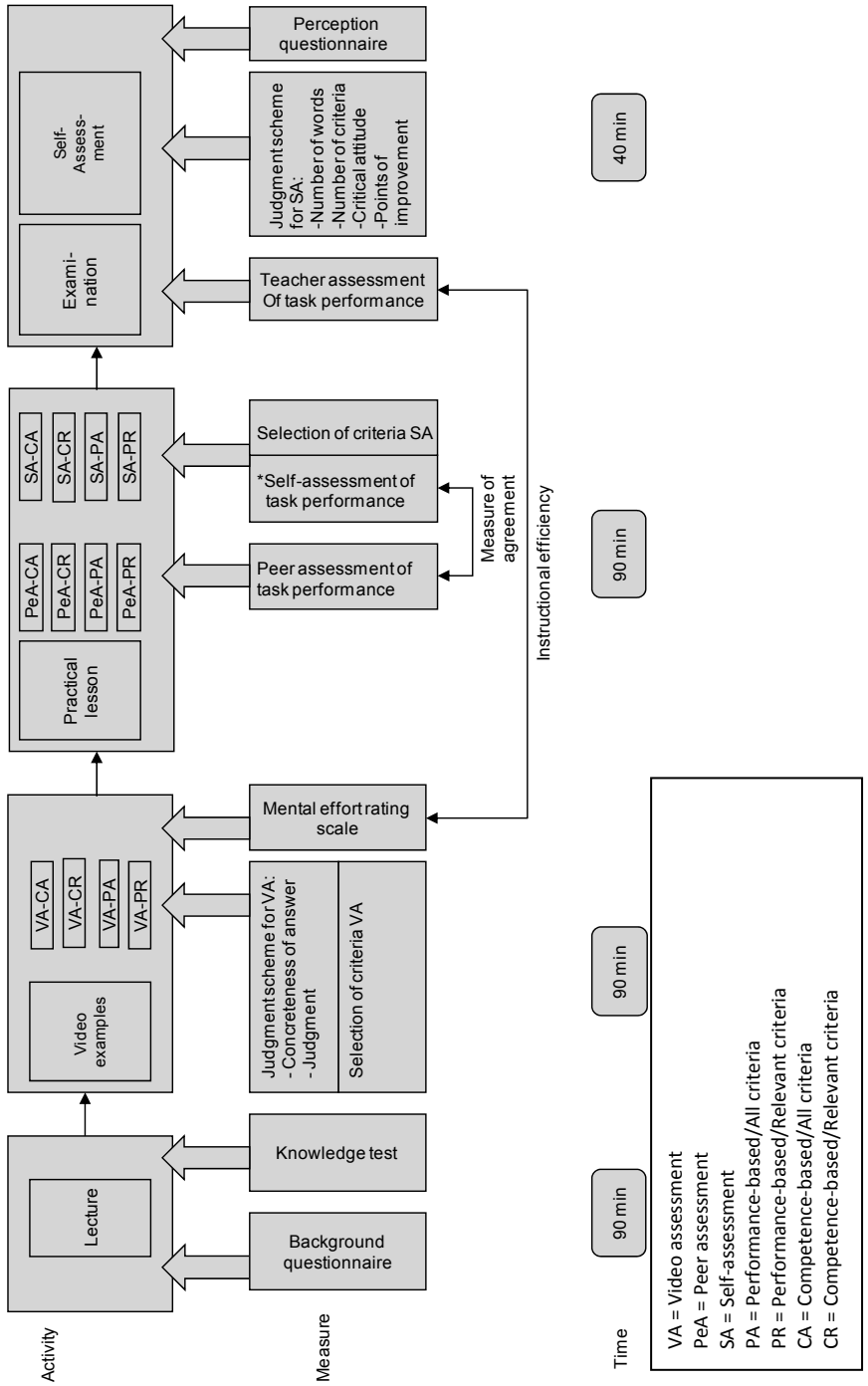

Figure 5.2: Overview of materials and measures.

Lecture

A ninety-minute lecture about the theoretical background of 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 the lectures. The researchers examined the content and the delivery of the lectures given by different teachers and observed no differences between them.

Video examples and video assessment

An electronic learning environment was developed which 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 method for novice learners to learn a new task (Stark, Kopp, & Fischer, 2009). The six fragments presented the consecutive components of the whole task of stoma care: (1) introduction, (2) preparation, (3) removing the old stoma bag, (4) fixing 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 tape, and they could watch each fragment a maximum of three times. On average, students watched the video fragments 1.19 times (*SD* = .22), and ANOVA showed no significant differences between the conditions in this regard.

In order to help students to make the assessment criteria more concrete, the students were asked to describe the nurse's behaviour after watching each video fragment. They were given an electronic list of preset 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 lists of assessment criteria were different for the four conditions and the four groups of students used the criteria relevant for their condition during all parts of the study. In the competence-based/all criteria condition, the assessment criteria were formulated as stoma care competences (VA-CA). The students were familiar with this type of criteria list, because it is routinely used during their training. Figure 5.3 shows examples of the criteria. The students received a list of all potentially relevant criteria and were instructed to select the criteria they considered relevant for assessing the video fragments.

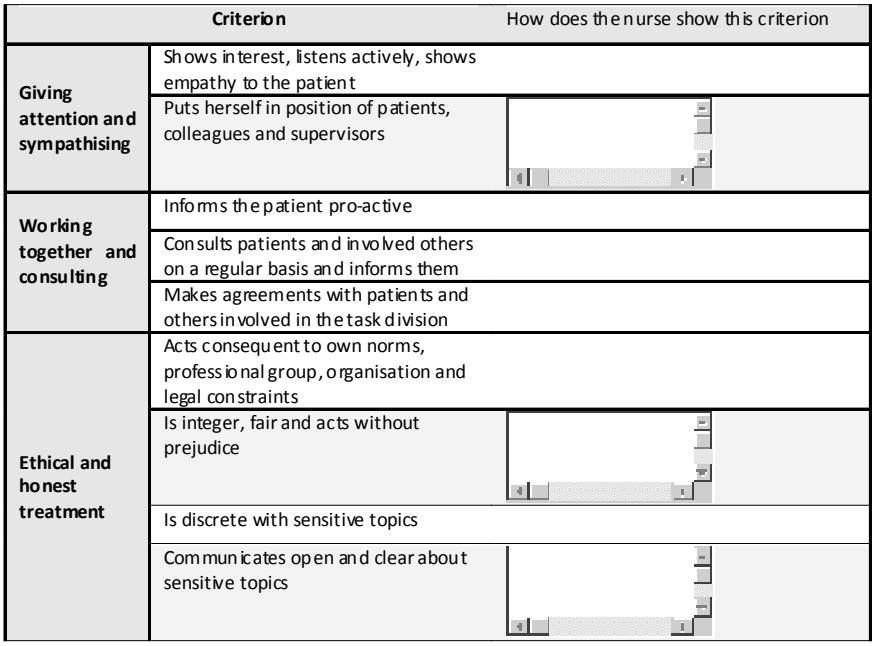

Figure 5.3: Screen dump of competence-based assessment criteria.

In the competence-based/relevant criteria condition, the students received 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 these criteria only (VA-CR).

In the performance-based/all criteria condition, the assessment criteria were formulated as context-specific stoma care skills (VA-PA). Examples are given in Figure 5.4. The students received a list of all potentially relevant criteria and were instructed to select the criteria they considered relevant for assessing the video fragment.

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

Figure 5.4: Screen dump of performance-based assessment criteria.

The students in the performance-based/relevant criteria condition received 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 these criteria only (VA-PR).

The students had ninety minutes to complete the video assessment and spent on average 48.65 minutes (*SD* = 11.96) working in the electronic learning environment. ANOVA revealed a significant difference between the conditions, with assessment time varying from 56.35 minutes (*SD* = 17.41) for the competence-based/all criteria condition, 46.39 minutes (*SD* = 9.80) for the competence-based/relevant criteria condition, 46.74 minutes (*SD* = 6.70) for the performance-based/all criteria condition, to 45.25 minutes (*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 smallest amount of time on the learning task, which made the experiment more conservative.

Practice session, peer assessment, and self-assessment

A ninety-minute practice session was developed, in which students practiced 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 rele-

vant criteria using a four 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-asses their own task performance in this way. Students in the all criteria condition were asked to indicate which criteria they considered relevant for the task at hand.

Test and self-assessment
A test task was developed in which the students individually performed the task of stoma care with a simulated patient, i.e., an actor who played the patient role. 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 write their answer to the question: Assess your 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, Brand-Gruwel, and van Merriënboer (2006) and with proven reliability (Fastré, van der Klink, & van Merriënboer, 2010). Table 5.1 shows 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.

Table 5.1: 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 Dutch.

Knowledge test
After the lecture, a 15-item multiple choice test was administered to assess the students' knowledge about stoma care. 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: Did the student give a concrete description of the nurse's behaviour (e.g., "she reassures the patient by saying I understand that it's difficult for you to have someone else do this") (0 = no concrete answer, 1 = concrete answer) and Did the student give 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 all) 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, Kester, and van Merriënboer (2009). The scale points ranged from very low mental effort (1) to very high mental effort (7). 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 four-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.

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 attended a four-hour training session on performance assessment one week before the actual assessment in order to enhance interrater agreement. Each student was assessed by a different pair of two randomly selected teachers. The average correlation between the scores of all pairs of teachers was $r = .53$.

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), students' critical attitude toward their own performance (0 = no critical attitude, 1 = critical attitude), and the presence of points of improvement (0 = no points of improvement, 1 = points of 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

The students evaluated their learning experience by rating the following aspects of perception of the learning environment on a four-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, Kanselaar, van der Linden, & Lodewijks, 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, Eghrari, Patrick, & Leone (1994), translated into Dutch by Martens and Kirschner (2004). Table 5.2 shows the Cronbach's alpha scores of the perception scales; internal consistencies ranged from .67 to .89, which is considered acceptable to high.

Table 5.2: 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 would like to conduct more learning tasks because they are useful.

[a] Items have been translated from Dutch.

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 & van Merriënboer, 1993; van Gog & Paas, 2008). Performance and mental effort scores were standardized, and the *z*-scores were entered into the formula:

$$E = \frac{Z_{Performance} - Z_{MentalEffort}}{\sqrt{2}}$$

A combination of relatively low mental effort with relatively high test task performance was indicative of high instructional efficiency, and 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 and they remained in the same group for the duration of the study. The groups first assessed the video fragments (maximum time ninety minutes) and immediately after that took part in the practice

session followed by peer and self-assessments, which together took another ninety minutes.

One week after the practice session, the students took the test and assessed their own performance (forty minutes). Student test task performance was subsequently assessed by two teachers. The evaluation questionnaire was administered immediately after the test.

Results

The results on the dependent variables in the test phase and the learning phase, and the students' perceptions are reported consecutively. Two-way ANOVA was conducted to test for effects of Relevance of criteria (only relevant versus all potentially relevant criteria), Type of criteria (performance-based versus competence-based), and their interaction. For all analyses, the significance level was set to .05. Partial Eta-squared is given as a measure of effect size, with η_p^2 = .01 indicating a small effect, η_p^2 = .06 a medium effect, and η_p^2 = .14 a large effect.

Test Phase

Table 5.3 presents the means and standard deviations for the dependent variables in the test phase.

Table 5.3: Means and standard deviations for dependent variables in the test phase.

	Competence-based / All criteria (*n* = 23)		Competence-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

A main effect of Relevance (F(1, 89) = 3.178, MSE = .022, p = .04, η_p^2 = .034) was found for test task performance as assessed by the 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.

A main effect of Type of criteria (F(1, 89) = 9.483, MSE = 8.000, p = .00, η_p^2 = .096) was found for instructional efficiency, indicating higher efficiency for the performance-based groups (M = .29, SD = 1.01) than for the competence-based groups (M = -.30, SD = .81). A marginally significant effect of Relevance was found (F(1, 89) = 1.817, MSE = 1.533, p = .09, η_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 5.4 presents the means and the standard deviations for the dependent variables in the learning phase.

Table 5.4: Means and standard deviations for dependent variables in the learning phase.

	Competence-based / All criteria (*n* = 23)		Competence-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	-	-

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 ($M = -.54$, $SD = 1.38$). Also a main effect of Type of criteria was 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 competence-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 competence-based groups (*M* = .67, *SD* = .47). 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 ($F(1, 89) = 3.130$, $MSE = .538$, $p = .04$, $\eta_p^2 = .034$). Visual inspection of the interaction (Figure 5.5) shows that the performance-based/relevant criteria group gave more concrete answers than the other groups, which all 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.

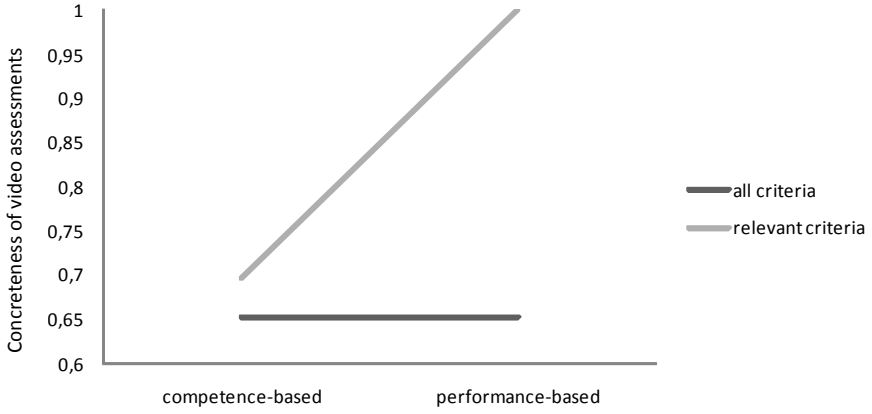

Figure 5.5: Difference in concreteness of video assessments.

Selection of criteria during the video assessments was only analyzed 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 ($t(44) = 17.68$, $p = .00$), indicating that the performance-based group selected a much higher number of relevant criteria (*M* = 34.61, *SD* = 3.74) compared to the competence-based group (*M* = 9.91, *SD* = 5.56). A significant difference was also found for the selection of irrelevant criteria ($t(44) = -2.876$, $p = .01$), with a much lower number of irrelevant criteria selected by the performance-based group (*M* = 15.09, *SD* = 11.75) compared to the competence-based group (*M* = 49.39, *SD* = 55.98).

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 competence-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 revealed a main effect of Type of criteria (F(1, 86) = 2.753, MSE = .603, p = .05, η_p^2 = .031), indicating a higher score in the performance-based groups (M = 3.45, SD = .53) compared to the competence-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, η_p^2 = .037), indicating a higher score in the performance-based groups (M = 3.33, SD = .53) compared to the competence-based groups (M = 3.12, SD = .56). A moderate agreement between peer and self-assessment was found, r = .50, p = .00, indicating congruity of self- and peer assessments.

The selection of relevant and irrelevant criteria during self-assessment in the practice session was only analyzed in the all criteria groups. A t-test revealed a significant difference for relevant criteria (t(42) = 3.793, p = .00), indicating that students in the performance-based group selected more relevant criteria (M = 31.50, SD = 14.16) compared to the competence-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 5.5 shows the means and standard deviations for all scales.

Table 5.5: Means and standard deviations for evaluation questionnaire.

	Competence-based / All criteria (n = 23)		Competence-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

On task orientation, a marginal effect of Relevance was found (F(1, 85) = 3.031, MSE = .831, p = .09, η_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). However, a marginal interaction effect for task orientation shows that the positive effect of providing all criteria is limited to the competence-based groups (F(1, 85) = 2.967, MSE = .813, p = .09, η_p^2 = .034). Visual inspection of the interaction effect (Figure 5.6) reveals that the competence-based/all criteria group shows

higher perceived task orientation than the other three groups, which have roughly similar scores. Thus, only students in the competence-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.

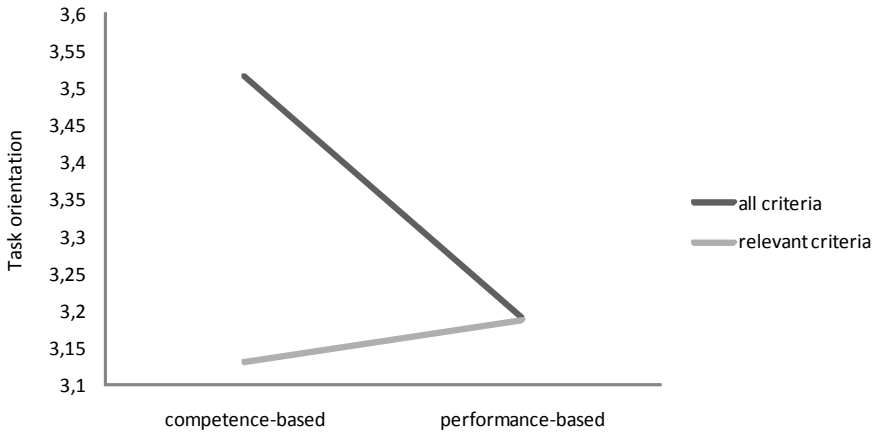

Figure 5.6: Difference in perceived task orientation.

Discussion

The goal of this study was to investigate the effects of type and relevance of assessment criteria in relation to students' test task performance and assessment skills. The first hypothesis, stating that students who receive performance-based criteria will show higher task performance, make better assessments, and invest less mental effort in the assessment than students who receive competence-based criteria is largely confirmed by the results. Students who received performance-based assessment criteria did indeed show higher task performance during the practice session as assessed by themselves and their peers, but test task performance as assessed by the teachers did not show this difference. Compared to the students in the competence-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 who invested less mental effort in the video assessments but achieved 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, Kester, Kirschner, & Bastiaens, 2008). The students who used performance-based criteria performed better during the practice session, during which they could consult the list of performance-based criteria, providing guidance as to what was important in the assessment. 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 practice the assessment with recourse to the criteria list may have been too short for the students to retain the criteria and apply them to improve task performance. In earlier research (Fastré, van der Klink, & van Merriënboer, 2010), the provision of performance-based criteria actually improved test task performance. Although this finding was not replicated in the current 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 & Paas, 2008).

The second hypothesis, stating that students who receive only the relevant criteria will show higher task performance and make better assessments than students who receive all criteria is also largely confirmed by the results. As expected, relevant criteria were associated with higher test task performance. Students who received 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, van der Klink, Sluijsmans, et al. (2010). The absence of an effect of relevance of criteria on mental effort suggests that cognitive load may be of a different nature depending on relevance of criteria (Sweller et al., 1998). For the all criteria groups, the invested mental effort may have been 'extraneous', i.e., ineffective for learning, with students engaging in unsuccessfully efforts to distinguish between relevant and irrelevant criteria. By contrast, the invested mental effort may have been effective, i.e., '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.

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 course, to task orientation, in particular, with higher scores of the competence-based/all criteria group. This is probably due to the strong similarity of the

criteria given to this group with the assessment instrument they were familiar with in their regular educational program, such as the list of 25 general competences. 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 relevance of criteria and the type of cognitive load (germane or extraneous) that students experience. As a related issue, effects of relevance of criteria should be examined among more advanced students. As students progress through a program, they are likely to become better able to translate competences into desired performances, which would reduce the cognitive load associated with competence-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. As competence development is a process that requires substantial training (van Merriënboer & Kirschner, 2007), the limited duration may have precluded the achievement of especially the higher levels of the complex skill of self-assessment. The self-assessment format, with one open question, may also have been suboptimal, because this type of question may have invited short and superficial responses, making it difficult to find differences between the groups.

This study was conducted among students from three different institutes for secondary vocational education. We therefore think that the findings can be generalized to a broader population. The results show that novice students should preferably be provided with relevant, performance-based assessment criteria in order to improve their assessment skills and task performance. In today's practice, however, novice students are provided with one list of 25 general competences, which may be a barrier to effective learning, because it does not help them to learn what acceptable task performance looks like. It should be noted that formulating performance-based criteria and determining their relevance for each learning task is a time-consuming process. Teachers need training and further support in formulating and using assessment criteria. Furthermore, more advanced students must learn to connect competences to performance and to distinguish between relevant and irrelevant competences, because these skills are indispensable for their professional development.

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

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*, 439-454.

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*, 523-538.

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.

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.

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

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*, 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*, 301-313.

Fastré, G. M. J., van der Klink, M. R., and 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,* 517-532.

Fastré, G. M. J., van der Klink, M. R., Sluijsmans, D., & van Merriënboer, J. J. G. (2010). *Drawing students' attention to relevant assessment criteria: Effects on self-assessment skills and performance.* Manuscript submitted for publication.

Grégoire, J. (1997). Diagnostic assessment of learning disabilities: From assessment of performance to assessment of competence. *European Journal of Psychological Assessment, 13,* 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*, 172-186.

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.

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.

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*, 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, November). 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 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.

CHAPTER 6
General discussion

This chapter presents an overview of the main findings and conclusions of the literature review and the empirical studies presented in this dissertation. Next, limitations and suggestions for future research are discussed. Furthermore, practical implications and a final conclusion are provided.

Main findings and conclusions

In order to deal with the challenging demands of their future workplaces, students need to develop self-directed learning skills which help them to become lifelong learners. In particular, they must develop sustainable assessment skills that enable them to recognize their own learning needs and to translate these learning needs into future actions (Boud, 2000; Boud & Falchikov, 2006). The introduction of competence-based educational approaches in secondary vocational education also places emphasis on changing assessment needs. The question is how assessment in vocational education should be designed to prepare students to become lifelong learners.

A literature review (Chapter 2) was conducted to investigate the nature of sustainable assessment skills, which enable students to assess their own performance and to continue learning throughout life. A model for developing sustainable assessment skills was presented consisting of three components: (a) conditions necessary for developing sustainable assessment skills, (b) constituent parts of sustainable assessment skills, and (c) educational methods for guiding students in the development of sustainable assessment skills. The three empirical studies following the literature review investigate the role of assessment criteria in the development of sustainable assessment skills, addressing the main question of this dissertation: *What kind of assessment criteria do novice students in the domain of Nursing and Care need in order to develop sustainable assessment skills and to become competent professionals?*

Three experimental studies were conducted in which Nursing & Care students were provided with: (1) only relevant versus all possibly relevant assessment criteria in Study 1 (Chapter 3), (2) performance-based versus competence-based assessment criteria in Study 2 (Chapter 4), and (3) a combination of these two in Study 3 (Chapter 5). The effects of these conditions on the development of (a) sustainable assessment skills, (b) task performance, (c) mental effort and instructional efficiency, and (d) students' perceptions of the learning environment will be summarized in the next sections.

Sustainable assessment skills

The effects of providing students with relevant criteria, performance-based criteria, and a combination of both were measured for three different sub skills of sustainable assessment: (a) the accuracy of self-assessments (Study 1), (b) the quality of self-assessments (Studies 2 and 3), and (c) the quality of assessments of others' performance (Studies 2 and 3).

Accuracy of self-assessment

In the first empirical study the accuracy of students' self-assessment skills was measured by the proportion of agreement between teacher assessments and students' self-assessments. Results reveal no differences in self-assessment accuracy between students in the different conditions.

Quality of self-assessment

To gain further insight in the self-assessment process of students, the focus of Studies 2 and 3 was on the quality of self-assessments. Regarding the overall quality of students' self-assessments, results reveal no differences between conditions. However, with regard to underlying skills, it can be concluded that providing students with performance-based assessment criteria helps them to differentiate between relevant and irrelevant criteria when they use the entire list with all criteria. Furthermore, results reveal that performance-based criteria have a positive, although marginal effect on the generation of points of improvement. Because the acquisition of complex skills such as self-assessment requires substantial practice (van Merriënboer & Kirschner, 2007), the limited duration of the studies may have yielded a level of self-assessment skills not high enough to find clear differences between the conditions. Overall, results show that providing students with performance-based assessment criteria seems the best starting point to improve the quality of their self-assessment skills.

Quality of assessment of other's performance

Assessment of the performance of an expert nurse shown in a video was part of Studies 2 and 3. The findings of these studies indicate that the quality of students' assessment of the performance of others is higher when they receive only relevant and performance-based criteria. Furthermore, providing students with only relevant assessment criteria leads to more concrete descriptions of the expert nurse's performance, especially when the criteria are performance-based. Providing only the relevant criteria also helps students to formulate a judgment of the expert nurse's performance. Furthermore, students had to assess the performance of their peers' performance in Studies 2 and 3, but the quality of this assessment was not investigated. Concluding, providing students with relevant performance-based assessment criteria leads to better assessments of the performance of others.

Task performance

The effects of providing students with only relevant and performance-based assessment criteria on task performance were measured. Studies 1 and 3 show that the use of relevant assessment criteria results in higher task performance. The findings of Study 2 indicate that performance-based criteria are beneficial for task per-

formance. It can be concluded that supporting students with relevant performance-based assessment criteria is most beneficial for students' task performance. Providing students with a long list of competence-based criteria, which is common practice, leads to significantly lower task performance and is detrimental to becoming a competent professional.

Mental effort and instructional efficiency

Students' invested mental effort while performing the assessment tasks was measured in the three studies. The findings of Study 1 reveal that providing students with the relevant assessment criteria leads to a higher invested mental effort, possibly because these criteria helped them to focus their attention on what is important in performance. In this case, the mental effort is positive for learning as it helps them to focus their attention. Results of Studies 2 and 3 show that providing students with performance-based assessment criteria leads to a lower investment of mental effort during assessment, probably because an interpretation of general competences in terms of concrete performance is no longer necessary. This means that most likely they do not have to invest mental effort that hinders their learning process. However, further research should shed light on the kind of mental effort that was involved in these studies. It is concluded that a reduction of mental effort during assessment can be accomplished by providing students with performance-based assessment criteria.

Additionally, in Studies 2 and 3 the instructional efficiency was measured by combining task performance with invested mental effort during assessment. Results show that performance-based assessment criteria lead to a higher instructional efficiency.

Student perceptions

Student perceptions of the learning environment were measured in all three studies. It appeared that, overall, students were positive about their learning environment. More specifically, in Study 1 an intriguing finding is that students in the all criteria group perceive the self-assessments as somewhat more relevant than students in the relevant criteria group. A possible explanation is that these students have to decide which criteria are relevant and which are not. In Study 3 an interesting marginal interaction effect on task orientation was found. The students who worked with all competence-based criteria indicated to have the clearest goals for the task at hand compared to the other group. An explanation might be that these students received an assessment instrument resembling the kind of instrument they were used to (i.e., a long list with competence-based criteria); while the other

groups were confronted with an instrument they were not familiar with. In general, it can be concluded that students were positive about the learning environment.

Theoretical implications

The empirical studies presented in this dissertation demonstrate that providing students with relevant performance-based assessment criteria improves their sustainable assessment skills. In particular, performance-based criteria appear to offer students support in the initial stage of acquiring and improving self-assessment skills. The presentation of only relevant performance-based assessment criteria contributes to students' task performance. In addition, performance-based criteria require less mental effort to achieve the same or even higher task performance, leading to higher instructional efficiency.

These results contribute to the further specification of the model for developing sustainable assessment skills presented in Chapter 2. The studies investigated methods to guide students in their development of sustainable assessment skills and focused on the nature of effective assessment criteria. But all aspects of the model were to some degree involved in the studies and will be discussed below.

Conditions for developing sustainable assessment skills

The conditions necessary for developing sustainable assessment skills were taken as a starting point in the empirical studies: (a) constructive alignment, (b) active learners, and (c) transparency of assessment criteria.

Constructive alignment

Good constructive alignment between instruction and assessment was guaranteed by fully integrating assessment activities in the learning tasks. The alignment with long-term purposes was realised by showing students an expert nurse demonstrating the intended performance. This helps students to become aware of the long-term purpose of performing well in their future profession. In the future, constructive alignment could be further strengthened by letting students decide on their own strategies to learn the demonstrated behaviour.

Active learners

Students were active learners because they received frequent feedback from others (teachers, peers, and work field supervisors), which they also had to compare with their self-assessments. Students had to assess their own performance, the performance of their peers, and the performance of an expert nurse. They were thus con-

stantly challenged to think about exemplary performance in relation to their own performance.

Transparency of assessment criteria
The transparency of assessment criteria was central in the presented studies because it was investigated what kind of assessment criteria students need to develop sustainable assessment skills. For novice students, the results show that the provision of relevant, performance-based criteria results in higher task performance and better assessment skills compared to either the provision of the whole set of relevant and irrelevant criteria or the provision of competence-based criteria.

Constituent parts of sustainable assessment skills

In an ideal situation, students self assess their performance and learning by first identifying criteria (a) explicitly provided by others, (b) implicitly used by others who assess their performance, and (c) implicitly used by others who demonstrate intended performance. One of the goals of the presented studies was to improve students' skills to self assess performance with a list of assessment criteria explicitly provided to them. The studies show that these criteria have to be performance-based and relevant for the task at hand. Students were not challenged to identify criteria implicitly used by others, although they could infer these criteria from the behaviour of the expert nurse and their peers. Yet, they were given guidance because all assessment criteria that were -possibly- relevant for that particular performance were made explicit to them. Moreover, the studies focused on improving the immediate task performance without highlighting the importance of learning and performance improvement over a series of learning tasks.

A second step in assessing performance and learning is to identify discrepancies (a) between own performance and learning and previously identified criteria, and (b) between self-assessments of performance and learning and assessments made by others. Indifferent of the specific assessment criteria provided to students, students had to self assess performance by identifying discrepancies between their own performance and given assessment criteria in all studies. They were better able to assess their own performance when the assessment criteria were performance-based and limited to the relevant criteria. Students were not explicitly challenged to identify discrepancies between their own assessments and the assessments made by others (teachers, peers, supervisors).

Based on a profound assessment of their current performance and their performance improvement over a series of learning tasks, students must generate goals for future performance and learning as a next step in their development of sustainable assessment skills. The generation of goals for future performance was included in the presented studies because students were challenged to formulate

points of improvement based on their self-assessments. Students were better able to formulate points of improvement when the assessment criteria were performance-based. However, they were not explicitly stimulated to set new learning goals for the long term.

Methods for guiding students in the development of sustainable assessment skills

The focus of this dissertation was on instructional methods for the development of sustainable assessment skills, especially for the development of skills necessary to identify relevant assessment criteria.

A first method to help students develop sustainable assessment skills is giving them examples of expert performance (i.e., modelling; van Merriënboer, Jelsma, & Paas, 1992). The presented studies used video-based modelling examples of expert performance that helped students to derive relevant assessment criteria. The aim was not to investigate whether offering students modelling examples had an effect on self-assessment skills, but students were challenged to analyse the examples from a particular perspective because they received different types of criteria. It may be concluded that students are better in analyzing modelling examples when they are provided with only relevant rather than all possibly relevant performance-based criteria. Compared to conventional instruction, students were very enthusiastic about the use of video-based modelling examples.

A second method to help students develop sustainable assessment skills is offering them assessment exercises in which they use assessment criteria to judge performance of experts and other students as well as their own performance. In the presented studies, students were constantly challenged to think about good performance. In accordance with the suggestions made by Sadler (1989), our results show that students should receive sufficient practice with given relevant criteria in the beginning of their study. Furthermore, results show that these assessment criteria are preferably performance-based to develop the necessary assessment skills.

Future research and limitations

The results of the empirical studies presented in this dissertation contribute to the further specification of the model for developing sustainable assessment skills. However, some elements of the model were not empirically tested in the studies described in this dissertation. For further refinement of the model, research should focus on at least the following aspects. First, to fully develop sustainable assessment skills, further investigation is necessary on alternative methods to guide students in the development of these skills. The studies in this dissertation focused on guiding students in the development of the skills necessary to identify relevant assessment

criteria without paying attention to the subsequent sustainable assessment skills, that is, to compare and analyze assessments of own performance and learning, and to generate goals for future performance and learning.

Second, in this dissertation, the model of sustainable assessment skills was further specified for novice students only. Future research should explore whether advanced students have different needs concerning the development of sustainable assessment skills. It should be examined whether advanced students are capable to independently select relevant assessment criteria or can be taught to acquire this skill, because it is an important part of sustainable assessment. Furthermore, research should shed light on the gradual transition from providing students with performance-based assessment criteria to increasingly more competence-based criteria.

In addition to further research on the elements of the model not studied in this dissertation, future research should also tackle the limitations of the reported studies. A first limitation is that the comparison of self-assessments with other assessments mainly took place at the closure of the experiment, when the teachers informed their students on assessment results. However, this comparison contains useful information for students that might help them to gain further expertise in assessment and to improve their performance. In future research, it should be explored whether explicitly letting students compare their self-assessments with other assessments of performance helps them to become competent professionals.

A second limitation is that, in all three studies, the focus of assessment was on the task level instead of the learning process; performance improvement over learning tasks was not explicitly taken into account. The assessment of task performance is a good starting point for assessing the learning process but in future research specific attention should be given to improving assessment of the learning process because this is also an important sustainable assessment skill.

The relatively short duration of the studies is a third limitation. Future research should examine long-term interventions to explore the development of sustainable assessment skills over time. As the development of highly complex skills, like self-assessment, is a time-consuming process, long-term interventions may provide more insight in how instructional methods best support the different phases in the development process.

The studies reported in this dissertation were conducted in three different institutes for secondary vocational education, making it plausible that the results can be generalized within the field of vocational education. However, future research is necessary to find out if the results can indeed be generalised to other domains and other contexts. For example, can similar problems be expected and can the same rules be applied for a study program in car mechanics or in electrotechnics? And can the results be generalised to contexts outside vocational education, for example to

pre-university education or post-graduate university programs? These are interesting topics for further investigation.

Practical implications and conclusion

Three practical implications arise from this dissertation. First, with respect to assessment criteria, it is important to provide novice students with task-relevant, performance-based assessment criteria. On the one hand, such criteria help students to become familiar with desired task performance and, on the other hand, they support them in the initial development of sustainable assessment skills. When students are not supported in the selection of relevant assessment criteria, they will most likely select criteria they are already familiar with or which they like rather than the most appropriate ones. When students are provided with competence-based criteria instead of performance-based criteria, they have difficulties to recognize and show good task performance, because the competence-based criteria give them insufficient guidelines to demonstrate desired behaviour. Providing students with a list of 25 generic competences, as is often done in today's secondary vocational education in the Netherlands, is harmful for students' learning processes. As students progress through their study, it is important to stimulate a gradual transition from thinking in terms of observable behaviours to thinking in terms of underlying competences as students need to assess competences in their future professional life.

Second, with regard to teachers involved in competence-based education, it should be noted that it is a time-consuming process to formulate performance-based assessment criteria for each learning task. Training teachers in how to formulate and use performance-based assessment criteria is a prerequisite for assessments to be fruitful in competence-based education. Furthermore, teachers should be trained in gradually shifting assessment responsibilities from themselves to the students and in stimulating students to acquire increasingly more experience with competence-based criteria as they progress throughout the curriculum.

To conclude, the studies reported in this dissertation yield very promising results for the development of sustainable assessment skills in vocational education. In addition, it appears that this development can be done in a relatively simple and inexpensive way. Providing students with task-relevant performance-based assessment criteria has positive effects on both their task performance and their development of sustainable assessment skills.

References

Boud, D. (2000). Sustainable assessment: Rethinking assessment for the learning society. *Studies in Continuing Education, 22*, 151-167.

Boud, D., & Falchikov, N. (2006). Aligning assessment with long-term learning. *Assessment and Evaluation in Higher Education, 31*, 399-413.

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

Van Merriënboer, J. J. G., Jelsma, O., & Paas, F. G. W. C. (1992). Training for reflective expertise: A four-component instructional design model for complex cognitive skills. *Educational Technology Research and Development, 40*, 23-43.

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

Summary

Chapter 1 introduces the main research question of this dissertation: What kind of assessment criteria do novice students in Nursing and Care need in order to develop sustainable assessment skills and to become competent professionals? Sustainable assessment skills are the skills to self assess performance and learning as well as the skills to set goals for future performance and learning. Graduated students are competent professionals when they are capable task performers and continue learning in their future professions. Students need to develop sustainable assessment skills to become independent and self-directed professionals who continue learning throughout life. One prominent skill in sustainable assessment is to identify relevant criteria for performance and learning. In secondary vocational education such criteria are often formulated on a holistic level, not describing the concrete performance students have to show. In addition, students are often expected to independently select the appropriate criteria for the learning tasks they work on. The studies in this dissertation examine what kind of assessment criteria novice students in the domain of Nursing and Care need in order to develop sustainable assessment skills, and to become competent professionals.

Chapter 2 describes the theoretical framework on which the studies presented in Chapters 3-5 are based. It presents a model for developing sustainable assessment skills. The model combines different theoretical perspectives and consists of three components: (a) conditions necessary for developing sustainable assessment skills, (b) constituent parts of sustainable assessment skills, and (c) educational methods to guide students in their development of sustainable assessment skills. The three empirical studies following this chapter focus on the development of students' sustainable assessment skills. They emphasize the importance of transparent assessment criteria and study the effects of different kinds of criteria on students' task performance and self-assessment skills.

Chapter 3 describes an empirical study that examines the effects of providing first-year Nursing and Care students (N = 68) in secondary vocational education with exclusively relevant assessment criteria on their task performance and their self-

assessment skills. The study employs a 2 x 2 factorial design with the between-subjects factors Relevance (providing all possibly relevant criteria versus providing only the relevant criteria) and Program (nursing versus care). In the relevant criteria/nursing group (*n* = 18) and the relevant criteria/care group (*n* = 16) students' attention is drawn to the assessment criteria that are relevant for a particular learning task by highlighting them in a list with all possibly relevant criteria. They are compared to an all criteria/nursing group (*n* = 18) and an all criteria/care group (*n* = 6), in which students receive the list with all possible criteria for each task without highlighting the relevant ones. It is hypothesized that students who practice with only the relevant assessment criteria show higher task performance on a test task than students in the all criteria group. Concerning self-assessment skills, it is hypothesized that students who practice with the relevant assessment criteria are better in selecting relevant criteria and in avoiding the selection of irrelevant criteria in the test task, and will show more accurate judgements of their task performance than students who practice with all criteria. Finally, it is hypothesized that students who practice with the relevant criteria will be better able to formulate points of improvement than the other group. As an exploration, the invested mental effort of students who self assess performance with only the relevant criteria and with all criteria is compared. Furthermore, it is explored whether there are differences between nursing students and care students on task performance and self-assessment skills.

Results reveal that, in accordance with the hypothesis, students in the relevant criteria group outperform the students in the all criteria group on the test task. No significant effects are found on selecting relevant criteria and avoiding the selection of irrelevant criteria, and on the accuracy of students' self-assessment skills. Students in the relevant criteria group perceive a higher invested mental effort for conducting the assessment. Furthermore, nursing students outperform care students in the test task and care students select more criteria, relevant and irrelevant, in the test phase than nursing students. Care students in the relevant criteria group are better in generating points of improvement than care students in the all criteria group. All students are enthusiastic about the learning environment. As no results are found on students' self-assessment skills, in the next experimental study a further investigation of the quality of students' self-assessment skills is conducted.

Chapter 4 describes an experimental study in which the effects of two different types of criteria on second-year Nursing and Care students' (*N* = 39) task performance and self-assessment skills are investigated. Two conditions are compared in which students receive either performance-based assessment criteria (*n* = 19) or competence-based criteria (*n* = 20). In the performance-based criteria group, students are provided with a preset list of criteria describing what students should do for the task at hand. In the competence-based criteria group, students receive a preset list of criteria describing what students should be able to do. It is hypothe-

sized that students in the performance-based condition show higher task performance than students in the competence-based condition. Furthermore, students in the performance-based group are expected to experience a lower invested mental effort during the assessment, and to demonstrate higher quality self-assessments than students in the competence-based group.

Results confirm the hypothesis on the positive effect of performance-based assessment criteria on students' task performance. Next to that, providing students with performance-based assessment criteria leads to a lower invested mental effort during the assessment. When students receive the performance-based assessment criteria they show a higher test task performance with a lower investment of mental effort, leading to a higher instructional efficiency. The hypothesis regarding assessment skills is not entirely confirmed; however, students provided with performance-based criteria are better in formulating points of improvement. Finally, students do not differ in their perceptions of the learning environment. The findings indicate that both groups are positive about the learning task as a whole.

The experimental study described in Chapter 5 combines the previous two studies by investigating the effects of type of assessment criteria (performance-based versus competence-based), relevance of assessment criteria (only relevant criteria versus all possibly relevant criteria), and their interaction on students' task performance and acquired assessment skills. Students of a second-year Nursing and Care program (N = 93) from three different nursing schools participate in the study. They are randomly assigned to one of four conditions: competence-based/all criteria (n = 23), competence-based/relevant criteria (n = 23), performance-based/all criteria (n = 23), and performance-based/relevant criteria (n = 24). Three different phases were distinguished: (1) a video example where students assess an expert nurse's performance, (2) a practical lesson with peer and self-assessment, and (3) a test phase with teacher and self-assessment. It is hypothesized that students who receive performance-based criteria show higher test task performance, better assessment skills, and lower invested mental effort than students who receive competence-based criteria. Furthermore, it is hypothesized that students who receive only the relevant assessment criteria will show higher task performance and better assessment skills than students who receive all the assessment criteria. Additionally, interaction effects are explored between relevance and type of assessment criteria, for which it is hypothesized that the combination of relevant and performance-based criteria is most beneficial for learning.

Results largely confirm the first hypothesis. Students who receive performance-based assessment criteria score higher on task performance during the practical lesson as assessed by their peers and by self-assessments, although they do not show significantly higher task performance in the test phase as assessed by their teachers. Furthermore, students in the performance-based criteria groups show higher-quality video assessments, select more relevant criteria and less irrelevant

criteria during these video assessments, and select more relevant criteria during the self-assessments in the practical lesson than students in the competence-based criteria groups. Students in the performance-based criteria groups also invest lower mental effort during the video assessments, leading to a higher instructional efficiency because they reach the same test task performance as students in the competence-based criteria groups but with less effort. The second hypothesis is also largely confirmed. Students who receive only the relevant criteria show higher test task performance. They also show a higher quality of their video assessments and a more frequent use of judgments in their assessments, although they do not show better self-assessments in the test phase. Next to that, the students who receive only the relevant criteria show a somewhat higher instructional efficiency than students who received all criteria, indicating a more favourable ratio between invested mental effort and test task performance. An interaction effect is found on the concreteness of the assessments of the video. In line with our expectations, students in the performance-based/relevant criteria group give more concrete answers in the assessment of video examples than the other students. All students are positive about the learning task, but the students who receive all competence-based criteria indicate to have the clearest idea of the goals of the task, probably because their instruction most resembles their usual practice.

Finally, Chapter 6 presents an overview of the results and general conclusions of the studies described in Chapters 3-5. The general conclusions presented in this chapter relate to the effects of providing relevant and performance-based assessment criteria on students' task performance and self-assessment skills. It is concluded that providing students with the relevant performance-based assessment criteria is a necessary step in improving students' sustainable assessment skills. Especially their task performance, their assessment skills when assessing others, and their generation of points of improvement are positively influenced. Theoretical implications focus on the conditions and methods of the model for developing sustainable assessment skills. For novice students, one of the conditions is that they need relevant performance-based assessment criteria. These criteria help them to become better assessors of others' task performance. This leads to the second implication that students need an extensive amount of assessment exercises to develop their assessment skills. Future research should shed light on the further development of sustainable assessment skills. Chapter 6 concludes with some practical implications that follow from the research presented in this dissertation. It is important for novice students to be provided with the relevant performance-based assessment criteria but students as well as teachers need training in working with these criteria to become better assessors.

Nederlandse samenvatting

Het eerste hoofdstuk van dit proefschrift beschrijft de onderzoeksvraag: Welke beoordelingscriteria hebben beginnende studenten Verpleging en Verzorging nodig om duurzame beoordelingsvaardigheden te ontwikkelen en om competente professionals te worden? Duurzame beoordelingsvaardigheden zijn de vaardigheden om de eigen prestaties en het eigen leren te beoordelen en om doelen te stellen voor toekomstige prestaties en leren. Gediplomeerde studenten zijn bekwame professionals als ze competente taakuitvoerders zijn en blijven leren in hun toekomstig beroep. Studenten moeten duurzame beoordelingsvaardigheden ontwikkelen om onafhankelijke professionals te worden die in staat zijn een leven lang te leren. Een belangrijke vaardigheid in duurzaam beoordelen is het identificeren van de relevante criteria voor prestaties en leren. In het middelbaar beroepsonderwijs zijn deze criteria vaak geformuleerd op een algemeen niveau, waarbij het concrete gedrag dat studenten in een leertaak moeten tonen niet beschreven wordt. Daarnaast worden studenten vaak geacht om zelf de relevante criteria voor een bepaalde leertaak te selecteren. De studies in dit proefschrift onderzoeken welke beoordelingscriteria beginnende studenten Verpleging en Verzorging nodig hebben om duurzame beoordelingsvaardigheden te ontwikkelen en om competente professionals te worden.

Hoofdstuk 2 beschrijft het theoretisch kader waarop de studies in de hoofdstukken 3 tot en met 5 gebaseerd zijn. Het theoretisch kader wordt gepresenteerd in de vorm van een model voor het ontwikkelen van duurzame beoordelingsvaardigheden. Het model combineert verschillende theoretische perspectieven en bevat 3 componenten: (a) condities voor het ontwikkelen van duurzame beoordelingsvaardigheden, (b) onderliggende samenstelling van duurzame beoordelingsvaardigheden, en (c) onderwijskundige methodes om studenten te begeleiden in de ontwikkeling van duurzame beoordelingsvaardigheden. De drie empirische studies die volgen op het theoretisch kader richten zich op de ontwikkeling van de duurzame beoordelingsvaardigheden van studenten. De studies benadrukken het belang van

transparante beoordelingscriteria en bestuderen de effecten van verschillende soorten criteria op de prestaties en zelfbeoordelingsvaardigheden van studenten.

Hoofdstuk 3 beschrijft een empirische studie waarin het effect van het geven van uitsluitend relevante beoordelingscriteria op de prestaties en de zelfbeoordelingsvaardigheden van eerstejaars studenten Verpleging en Verzorging (*N* = 68) in het middelbaar beroepsonderwijs is onderzocht. De studie bestaat uit een 2 x 2 factorieel experiment met als factoren relevantie (het aanbieden van alle mogelijke criteria versus het aanbieden van enkel de relevante criteria) en programma (verpleging versus verzorging). In de relevante criteria/verpleging groep (*n* = 18) en de relevante criteria/verzorging groep (*n* = 16) wordt de aandacht van de studenten gericht op de relevante beoordelingscriteria voor een bepaalde leertaak door deze vet te drukken in een lijst met alle mogelijke relevante criteria. Zij worden vergeleken met een alle criteria/verpleging groep (*n* = 18) en een alle criteria/verzorging groep (*n* = 6), waarin de studenten de lijst met alle mogelijke criteria voor een bepaalde leertaak krijgen zonder dat de relevante criteria vetgedrukt zijn. De veronderstelling is dat studenten die oefenen met de vetgedrukte relevante beoordelingscriteria beter presteren op een testtaak dan studenten in de alle criteria groep. Met betrekking tot hun zelfbeoordelingsvaardigheden wordt verondersteld dat studenten die oefenen met de relevante criteria beter zijn in het selecteren van de relevante criteria en het vermijden van de selectie van irrelevante criteria in de testtaak, waar ze wel alle criteria aangeboden krijgen, en dat ze een juister oordeel over hun prestaties hebben dan studenten die oefenen met de lijst met alle criteria. Tenslotte wordt verondersteld dat studenten die oefenen met de relevante criteria beter zijn in het formuleren van verbeterpunten dan de studenten in de andere groep. Voorts wordt de mentale inspanning vergeleken van de beide groepen studenten die zichzelf beoordelen met de relevante criteria of alle criteria. Verder wordt geëxploreerd of er verschillen zijn tussen studenten Verpleging en Verzorging op prestaties en zelfbeoordelingsvaardigheden.

Resultaten tonen aan dat, in overeenstemming met de hypothese, studenten in de relevante criteria groep beter presteren dan studenten in de alle criteria groep op de testtaak. Er werden geen significante resultaten gevonden op het selecteren van de relevante criteria en het vermijden van de selectie van irrelevante criteria, noch op de juistheid van de zelfbeoordeling. Studenten in de relevante criteria groep ervaren een hogere mentale inspanning om de beoordeling uit te voeren. Daarnaast blijkt dat studenten Verpleging beter presteren op de testtaak dan studenten Verzorging en dat studenten Verzorging meer criteria selecteren in de testfase dan studenten Verpleging, zowel relevant als irrelevant. Studenten Verzorging in de relevante criteria groep zijn beter in het formuleren van verbeterpunten dan studenten Verzorging in de alle criteria groep. Alle studenten zijn enthousiast over de leeromgeving. Omdat er geen resultaten gevonden zijn op de zelfbeoordelings-

vaardigheden van studenten werd in de volgende experimentele studie de kwaliteit van de zelfbeoordelingen verder onderzocht.

Hoofdstuk 4 beschrijft een experimentele studie waarin de effecten van twee verschillende types criteria op de prestaties en de zelfbeoordelingsvaardigheden van tweedejaars studenten Verpleging en Verzorging (*N* = 39) onderzocht zijn. Er zijn twee condities vergeleken waarbij studenten ofwel prestatiegerichte criteria ontvingen (*n* = 19), ofwel competentiegerichte criteria (*n* = 20). In de prestatiegerichte criteria groep krijgen de studenten een lijst met beoordelingscriteria die beschrijven wat de student moet doen voor de betreffende taak. In de competentiegerichte criteria groep ontvangen de studenten een lijst met beoordelingscriteria die beschrijven waartoe de student in staat moet zijn. Verondersteld wordt dat studenten in de prestatiegerichte criteria groep beter presteren op de taak dan studenten in de groep die de competentiegerichte criteria ontvangen. Daarnaast wordt verondersteld dat studenten in de prestatiegerichte groep een lagere mentale inspanning ervaren tijdens de beoordeling en dat ze een hogere kwaliteit van zelfbeoordeling vertonen dan studenten in de competentiegerichte groep.

Het positieve effect van prestatiegerichte beoordelingscriteria op de prestaties is in de onderzoeksresultaten waarneembaar. Daarnaast blijkt dat het aanbieden van de prestatiegerichte criteria leidt tot een lagere mentale inspanning tijdens de beoordeling. Studenten met de prestatiegerichte criteria presteren beter met een lagere mentale inspanning, wat leidt tot een hogere efficiëntie van de instructie. De hypothese betreffende de beoordelingsvaardigheden wordt niet volledig bevestigd, maar prestatiegerichte criteria dragen wel bij tot het beter formuleren van verbeterpunten. Tot slot blijkt dat studenten niet verschillen in hun percepties van de leeromgeving. De resultaten tonen aan dat beide groepen positief zijn over de leertaak in zijn geheel.

De experimentele studie, beschreven in hoofdstuk 5, combineert de vorige twee studies door te onderzoeken wat het effect is van het type beoordelingscriteria (prestatiegericht versus competentiegericht), relevantie van beoordelingcriteria (enkel relevante criteria versus alle mogelijke relevante criteria), en hun interacties op de prestaties en beoordelingsvaardigheden van studenten. Tweedejaars studenten Verpleging en Verzorging (*N* = 93) uit drie verschillende scholen namen deel aan de studie. Zij zijn willekeurig toegewezen aan één van de vier condities: competentiegericht/alle criteria (*n* = 23), competentiegericht/relevante criteria (*n* = 23), prestatiegericht/alle criteria (*n* = 23) en prestatiegericht/relevante criteria (*n* = 24). Er werden drie verschillende fasen onderscheiden: (1) een videovoorbeeld waarbij de student de prestaties van een expertverpleegkundige beoordeelt, (2) een praktische les met zelfbeoordeling en beoordeling van de medestudenten, en (3) een testfase met docent- en zelfbeoordeling. De verwachting is dat studenten die de prestatiegerichte criteria krijgen beter presteren op de testtaak in de testfase, beter kunnen beoordelen, en minder inspanning hoeven te leveren voor deze beoordeling dan

studenten die de competentiegerichte criteria krijgen. Verder wordt verondersteld dat studenten die enkel de relevante criteria krijgen beter presteren en beter beoordelen dan studenten die alle criteria krijgen. Tevens is nagegaan of er sprake is van interactie-effecten tussen relevantie en type beoordelingscriteria waarbij verondersteld wordt dat de combinatie van relevante en prestatiegerichte criteria het meest voordelig is voor het leren.

De resultaten bevestigen grotendeels de eerste hypothese. Studenten die de prestatiegerichte beoordelingscriteria krijgen presteren beter in de praktische les zoals beoordeeld door hun medestudenten en door henzelf. Ze presteren echter niet beter in de testfase waar ze beoordeeld werden door hun docenten. Verder blijkt dat studenten in de groepen die werkten met de prestatiegerichte criteria beter waren in het beoordelen van de videofragmenten, meer relevante en minder irrelevante criteria selecteren bij die beoordeling en meer relevante criteria selecteren tijdens de zelfbeoordeling in de praktische les dan de studenten in de groepen die werkten met de competentiegerichte criteria. Studenten in de groepen die de prestatiegerichte criteria ontvingen ervaren ook minder mentale inspanning tijdens het beoordelen van de videofragmenten, wat leidt tot een hogere efficiëntie van de instructie, omdat ze dezelfde prestatie neerzetten als studenten in de competentiegerichte groepen maar met minder inspanning. De tweede hypothese is ook grotendeels bevestigd. Studenten die alleen de relevante criteria krijgen presteren beter op de testtaak. Ze vertonen een hogere kwaliteit van de videobeoordelingen en een frequenter gebruik van uitspraken over de kwaliteit van hun prestaties in hun beoordelingen, maar hun zelfbeoordelingen in de testfase zijn niet beter. Daarnaast is er in de groepen die enkel de relevante criteria ontvangen sprake van een iets hogere efficiëntie van de instructie wat leidt tot een meer positieve ratio tussen de mentale inspanning en de prestaties. Op de concreetheid van de videobeoordelingen werd een interactie-effect gevonden. In lijn met onze verwachtingen is de prestatiegerichte/relevante criteria groep in staat meer concrete antwoorden te geven bij het beoordelen van de videovoorbeelden dan de andere drie groepen. Alle studenten zijn positief over de leertaak, maar de studenten die alle competentiegerichte criteria krijgen geven aan een helder beeld te hebben van de doelen bij de taak, waarschijnlijk omdat hun instructie het meest hun dagelijkse praktijk benadert.

Tenslotte biedt hoofdstuk 6 een overzicht van de resultaten en algemene conclusies van de studies die zijn beschreven in hoofdstuk 3 tot en met 5. De algemene conclusies gaan over de effecten van het geven van de relevante en prestatiegerichte beoordelingscriteria op de prestaties en zelfbeoordelingsvaardigheden van studenten. De belangrijkste conclusie is dat het aanbieden van de relevante prestatiegerichte beoordelingscriteria een noodzakelijke stap is in het bevorderen van de duurzame beoordelingsvaardigheden van studenten. Vooral hun prestaties, hun beoordelingsvaardigheden bij het beoordelen van anderen en de formulering van

hun verbeterpunten worden door deze criteria positief beïnvloed. Theoretische implicaties focussen op de condities en de methodes van het model voor het ontwikkelen van duurzame beoordelingsvaardigheden. Voor beginnende studenten is één van de condities dat zij de relevante prestatiegerichte criteria nodig hebben om betere beoordelaars te worden. Dat leidt tot de tweede implicatie, namelijk dat studenten een aanzienlijke hoeveelheid oefening nodig hebben om hun beoordelingsvaardigheden te ontwikkelen. Toekomstig onderzoek moet zich richten op de verdere ontwikkeling van duurzame beoordelingsvaardigheden. Hoofdstuk 6 eindigt met enkele praktische implicaties van de studies gepresenteerd in dit proefschrift. Het is belangrijk voor beginnende studenten om met de relevante prestatiegerichte criteria te werken, maar zowel studenten als docenten moeten training krijgen in het werken met deze criteria om betere beoordelaars te worden.

ICO dissertation series

ico

In the ICO Dissertation Series dissertations are published of graduate students from faculties and institutes on educational research within the following universities: Eindhoven University of Technology, Leiden University, Maastricht University, Open University of the Netherlands, University of Amsterdam, University of Groningen, University of Twente, Utrecht University, VU University Amsterdam, and Wageningen University (and formerly Radboud University Nijmegen and Tilburg University).

167. Pijls, M. (13-06-2007). *Collaborative mathematical investigations with the computer: Learning materials and teacher help.* Amsterdam: University of Amsterdam.
168. Könings, K. (15-06-2007). *Student perspectives on education: Implications for instructional design.* Heerlen: Open University of the Netherlands.
169. Prangsma, M.E. (20-06-2007). *Multimodal representations in collaborative history learning.* Utrecht: Utrecht University.
170. Niemantsverdriet, S. (26-06-2007). *Learning from international internships: A reconstruction in the medical domain.* Maastricht: Maastricht University.
171. Van der Pol, J. (03-07-2007). *Facilitating online learning conversations: Exploring tool affordances in higher education.* Utrecht: Utrecht University.
172. Korobko, O.B. (07-09-2007). *Comparison of examination grades using item response theory: A case study*. Enschede: University of Twente.
173. Madih-Zadeh, H. (14-09-2007). *Knowledge construction and participation in an asynchronous computer-supported collaborative learning environment in higher education.* Wageningen: Wageningen University.
174. Budé, L.M. (05-10-2007). *On the improvement of students' conceptual understanding in statistics education.* Maastricht: Maastricht University.
175. Meirink, J.A. (15-11-2007). *Individual teacher learning in a context of collaboration in teams.* Leiden: Leiden University.
176. Niessen, T.J.H. (30-11-2007). *Emerging epistemologies: Making sense of teaching practices.* Maastricht: Maastricht University.
177. Wouters, P. (07-12-2007). *How to optimize cognitive load for learning from animated models.* Heerlen: Open University of the Netherlands.
178. Hoekstra, A. (19-12-2007). *Experienced teachers' informal learning in the workplace.* Utrecht: Utrecht University.
179. Munneke-de Vries, E.L. (11-01-2008). *Arguing to learn: Supporting interactive argumentation through computer-supported collaborative learning*. Utrecht: Utrecht University.
180. Nijveldt, M.J. (16-01-2008). *Validity in teacher assessment. An exploration of the judgement processes of assessors*. Leiden: Leiden University.

181. Jonker, H.G. (14-02-2008). *Concrete elaboration during knowledge acquisition*. Amsterdam: VU University Amsterdam.
182. Schuitema, J.A. (14-02-2008). *Talking about values. A dialogue approach to citizenship education as an integral part of history classes*. Amsterdam: University of Amsterdam.
183. Janssen, J.J.H.M. (14-03-2008). *Using visualizations to support collaboration and coordination during computer-supported collaborative learning*. Utrecht: Utrecht University.
184. Honingh, M.E. (17-04-2008). *Beroepsonderwijs tussen publiek en privaat: Een studie naar opvattingen en gedrag van docenten en middenmanagers in bekostigde en niet-bekostigde onderwijsinstellingen in het middelbaar beroepsonderwijs.* Amsterdam: University of Amsterdam.
185. Baartman, L.K.J. (24-04-2008). *Assessing the assessment: Development and use of quality criteria for competence assessment programmes*. Utrecht: Utrecht University.
186. Corbalan Perez, G. (25-04-2008). *Shared control over task selection: Helping students to select their own learning tasks*. Heerlen: Open University of the Netherlands.
187. Hendrikse, H.P. (22-05-2008). *Wiskundig actief: Het ondersteunen van onderzoekend leren in het wiskunde onderwijs*. Enschede: University of Twente.
188. Moonen, M.L.I. (26-09-2008). *Testing the multi-feature hypothesis: Tasks, mental actions and second language acquisition*. Utrecht: Utrecht University.
189. Hooreman, R.W. (18-11-2008). *Synchronous coaching of the trainee teacher: An experimental approach*. Eindhoven: Eindhoven University of Technology.
190. Bakker, M.E.J. (02-12-2008). *Design and evaluation of video portfolios: Reliability, generalizability, and validity of an authentic performance assessment for teachers.* Leiden: Leiden University.
191. Kicken, W. (12-12-2008). *Portfolio use in vocational education: Helping students to direct their learning.* Heerlen: Open University of the Netherlands.
192. Kollöffel, B.J. (18-12-2008). *Getting the picture: The role of external representations in simulation-based inquiry learning*. Enschede: University of Twente.
193. Walraven, A. (19-12-2008). *Becoming a critical websearcher: Effects of instruction to foster transfer.* Heerlen: Open University of the Netherlands.
194. Radstake, H. (14-05-2009).*Teaching in diversity: Teachers and pupils about tense situations in ethnically heterogeneous classes*. Amsterdam: University of Amsterdam.
195. Du Chatenier, E. (09-09-2009). *Open innovation competence: Towards a competence profile for inter-organizational collaboration in innovation teams*. Wageningen: Wageningen University.
196. Van Borkulo, S.P. (26-06-2009). *The assessment of learning outcomes of computer modelling in secondary science education*. Enschede: University of Twente.
197. Handelzalts, A. (17-09-2009). *Collaborative curriculum development in teacher design teams*. Enschede: University of Twente.
198. Nievelstein, F.E.R.M. (18-09-2009). *Learning law: Expertise differences and the effect of instructional support*. Heerlen: Open University of the Netherlands.
199. Visser-Wijnveen, G.J. (23-09-2009). *The research-teaching nexus in the humanities: Variations among academics*. Leiden: Leiden University.
200. Van der Rijst, R.M. (23-09-2009). *The research-teaching nexus in the sciences: Scientific research dispositions and teaching practice*. Leiden: Leiden University.
201. Mainhard, M.T. (25-09-2009). *Time consistency in teacher-class relationships*. Utrecht: Utrecht University.
202. Van Ewijk, R. (20-10-2009). *Empirical essays on education and health*. Amsterdam: University of Amsterdam.
203. Seezink, A. (18-11-2009). *Continuing teacher development for competence-based teaching*. Tilburg: Tilburg University.
204. Rohaan, E.J. (09-12-2009). *Testing teacher knowledge for technology teaching in primary schools*. Eindhoven: Eindhoven University of Technology.
205. Kirschner, F.C. (11-12-2009). *United brains for complex learning*. Heerlen: Open University of the Netherlands.
206. Wetzels, S.A.J. (18-12-2009). *Individualized strategies for prior knowledge activation*. Heerlen: Open University of the Netherlands.