Learning Environments Research (2006) 9: 45–66
DOI: 10.1007/s10984-005-9003-3

DOMINIQUE M. A. SLUIJSMANS, FRANS J. PRINS AND ROB L. MARTENS

THE DESIGN OF COMPETENCY-BASED PERFORMANCE ASSESSMENT IN E-LEARNING

Received 16 September 2004; accepted (in revised form) 12 September 2005

ABSTRACT. This article focuses on the design of competency-based performance assessment in e-learning. Though effort has been invested in designing powerful e-learning environments, relatively little attention has been paid to the design of valid and reliable assessments in such environments, leaving many questions to educational developers and teachers. As a solution to this problem, a systematic approach to designing performance assessments in e-learning contexts is presented, partly based on the 4C/ID model. This model enables the construction of realistic whole tasks instead of advocating education that is restricted to more isolated skills. A new assessment procedure also implies an alternative view of instructional design, learning and assessment. The requirements for the learning environment are addressed. Examples from a virtual seminar are presented to illustrate the design approach. The article concludes with the identification of possible pitfalls related to the approach and gives directions for future research.

KEY WORDS: computer-based learning, e-learning, learning environments, performance assessment

Institutions of higher education are confronted with a demand for competency-based learning (CBL), which is expected to narrow the gap between learning in the educational setting and future workplace performance (Bastiaens & Martens, 2000). In competency-based learning environments, developers try to confront learners with authentic, open problems and learning materials that have personal meaning for them and are presented in a variety of formats. Teaching methods are applied that arouse interest, activate prior knowledge, clarify meanings, and model appropriate learning strategies and reflective processes. Learners are supposed to use combinations of acquired skills and knowledge. The to-be-acquired knowledge and skills have to be integrated in educational activities, so that learners recognise a resemblance with tasks in the real world (Stoof, Martens, van Merriënboer, & Bastiaens, 2002). As in professional work contexts, more and more collaboration goes with computers (Strijbos, Kirschner, & Martens, 2004). Computers provide excellent opportunities for socio-constructivist learning experiences. "Emerging technologies of computer supported collaborative learning (CSCL) provide increasing opportunities for fostering learning in such an environment by creating on-line communities of learners. (. . .) It offers a dynamic collaborative environment in which learners can interact, engage in critical thinking, share ideas, defend and challenge each other's assumptions, reflect on the learning material,

ask questions, test their interpretations and synthesis, and revise and reconstruct their ideas" (Birenbaum, 2003, p. 21).

Many efforts have been made to implement CBL and assessment in face-to-face education, but it is still quite a struggle when it comes to the design of CBL and assessment in an electronic learning environment. In particular, the design of more performance-based assessment is a weak component in e-learning in that the emphasis is still much more on traditional testing than on assessment (Segers & Dierick, 2001). The Standards for Educational and Psychological Testing (American Educational Research Association, American Psychological Association & National Council on Measurement in Education, 1999, p. 3) define test as "an evaluative device or procedure in which a sample of a learner's behaviour in a specified domain is obtained and subsequently evaluated and scored using a standardized process." Testing is often a process of gathering data and returning results, instead of a process of providing opportunities for learning. Data from several assessments are used to make a decision about a learner's performance level.

Many authors signal a shift from a test culture to an assessment culture which strongly emphasises integration of instruction, learning and assessment (Biggs, 1996; Birenbaum, 2003). In contrast to traditional tests, we refer to assessments when they are based on multiple products or processes, such as essays, reflection papers, oral assessments, process analyses, group products, and work samples. The assessment task is described in terms of a certain performance that is perceived as worthwhile and relevant to the learner, and therefore can be defined as performance assessments (Wiggins, 1989). Performance assessment focuses on the ability to use combinations of acquired skills and knowledge, and therefore fits in well with the theory of powerful learning environments (Linn, Baker, & Dunbar, 1991). Because the goals as well as the methods of instruction are oriented towards integrated and complex curricular objectives, it is necessary for assessment practices to reflect this complexity and to use various kinds of assessments in which learners have to interpret, analyse and evaluate problems and explain their arguments. These assessments, which should be fully integrated in the learning process, provide information about learner progress and support learners in selecting appropriate learning tasks. The compatibility between instruction, learning, and assessment is described within the theory of constructive alignment (Biggs, 1996; see also Birenbaum, 2003). When there is alignment between what teachers want to teach, how they teach, and how they assess, teaching is likely to be more effective than when it is not. To pursue the theory of constructive alignment, it is worthwhile to invest in the design of performance assessments, because performance assessment provides multidimensional feedback for fostering learning (Birenbaum, 2003).

As stated before, the increased use of ICT plays an important role in the shift towards CBL and an assessment culture. ICT enables many forms of advanced socio-constructivist learning, such as CSCL, can increase the resemblance to professional contexts, allows simulations, and so on (Martens, 1998). Modern distance learning courses are often set up as realistic 'games' or simulations (e.g. Garris, Ahlers & Driskell, 2002). ICT here is intended to provide the 'virtual' reality as a motivating authentic problem and serves as a provider for CBL. It enables the construction of realistic 'whole tasks' with high authenticity. Despite these new educational possibilities, there is a strong emphasis on what is technologically possible, not on what is educationally desirable (van Merriënboer & Martens, 2002). This can be seen in the literature on testing in e-learning, which mainly focuses on tools that are item-based and that are directed at knowledge-based testing. A major disadvantage of such tests is that they tend to focus on the measurement of low-level retention of isolated facts rather than on the application of knowledge to solve ill-structured problems (Baker & Mayer, 1999; Reeves, 2000). Zhang, Khan, Gibbons and Ni (2001), for example, reviewed 10 web-based tools that were all based on the item type testing. One of the well-known applications of these item-based tools in e-learning is computerised adaptive testing (CAT). In CAT, the computer dynamically evaluates the ability of the student, resulting in a test that is adapted to each individual student. This fine-tuning is achieved by statistically adapting the test to the achievement level of each student while avoiding very easy or very difficult questions. Although Zhang et al. (2001) concluded that time and place independency were the main advantages of these tools, they also acknowledge that none of the tools make use of performance assessments. Web-based tests are yet far away from assessments that support relevant professional performance and learning.

Simulations of hands-on tasks are found useful for more *performance-based assessment* in e-learning (Shavelson, 1991). In the technical area, for example, simulators are developed for certifying pilot competencies (Bennett, 1999), or for training and assessing skills to operate submarine periscopes (Garris et al., 2002). Recent research projects focus on the assessment of problem-solving skills (e.g. Mayer, 2002; O'Neil, 2002). Here the computer keeps a record of every move made by the student in solving a task in order to provide a detailed profile of his or her performance for assessment. Video recording is also a realistic option for performance assessment (e.g. in teacher education contexts). A video recording can be applied for educational activities as analysis of the observation, peer review or other assignments.

In summary, to date, ICT is often used as a technology to simulate the context of professional practice in education. But we know little about

how to bring the assessment in e-learning in line with these often complex learning environments.

To address this problem, this article focuses on the design of competency-based performance assessment in e-learning and the implications of integrating performance assessment in e-learning. Because assessment that is strongly embedded in instructional practice in general, and in e-learning in particular, is very hard to find, the lack of structured frameworks to guide assessment design is understandable. This article presents design guidelines that support and guide the design of sound performance assessments in e-learning. Throughout the article, examples from a distributed case-based CSCL-course called the European Virtual Seminar (EVS), designed at the Open University of The Netherlands, are presented to illustrate the implementation of competency-based performance assessment in an e-learning environment, as well as possible problems that can be encountered during implementation of performance assessment activities in e-learning. In this EVS course, multidisciplinary student groups were asked to conduct research based on an authentic case concerning sustainable development and enlargement of the European Union. Students had to write a group report in which they had to integrate different disciplinary views on the problem described in the case, and provide recommendations to the European Union for a policy change. The student groups collaborated in Blackboard 5®, a virtual learning environment (VLE), and they could communicate by chat facilities and discussion boards. For a detailed description of the EVS course, refer to Prins, Sluijsmans, Kirschner and Strijbos (2005).

1. Designing Performance Assessments

For the alignment of instruction, learning, and assessment, learning tasks should be directly related to the performance assessment tasks at the end of a study unit. In contrast with the design procedure of most teachers, in which the learning tasks are designed prior to the assessment, Stiggins (1987) states that the design of the assessment should be the first step in educational design. For this, he formulated four general guidelines to design performance assessments (see Table I).

First, the *purpose* of the performance assessment has to be defined. Several important questions are in order (Herman, Aschbacher & Winters, 1992). What important cognitive skills or attributes do students have to develop? What social and affective skills or attributes do students have to develop? What metacognitive skills do students have to develop? What types of problems do they have to be able to solve? What concepts and principles do students have to be able to apply? This first step results in a skill decomposition in which the relevant skills are hierarchically ordered

TABLE I
A Step-by-Step Approach for Designing Performance Assessments

Step	What to do?
Define the purpose of the assessment	List the skills and knowledge that you wish to have students learn as a result of completing a task.
Define performance assessment tasks	Design a performance task which requires the students to demonstrate these skills and knowledge.
Define performance criteria	Develop explicit performance criteria which measure the extent to which students have mastered the skills and knowledge.
Create performance scoring rubrics	Use one scoring system or performance rubric for each performance task. The performance criteria consist of a set of score points which define in explicit terms the range of student performance.

(van Merriënboer, 1997). For example, in the case-based CSCL course European Virtual Seminar (EVS), the course designers and teachers described the purpose of the performance assessment as follows: "After having participated in the EVS course, students should be able to (1) describe the concept of sustainable development, (2) give an overview of the relation between EU-enlargement policy-related issue and sustainable development, (3) make a link between regional, national, and European issues concerning sustainable development, (4) work with students with different disciplinary and cultural backgrounds and exchange views on sustainable development in Europe, (5) participate effectively in a computer conference, and (6) use the computer conference for collaborative learning purposes." The first three purposes reflect the knowledge that students have to acquire and the latter three concern the skills that have to be learned as a result of completing the tasks in the EVS course. Also this e-learning course illustrates the frequently-stated wish of teachers of CSCL courses that students have to acquire and use collaborative learning skills. However, opportunities to deliberately practise and learn these skills are often limited. In the EVS course, students received supportive information concerning collaborative learning in small virtual groups before they started conducting the research about sustainable development.

When the purpose of the assessment is defined, decisions are made concerning the *performance assessment task*. The performance assessment task can be a product, behaviour or extended written response to a question that requires the student to apply critical thinking skills. Some examples of performance assessment tasks include written compositions, speeches, and research projects. It is important that the performance assessment task can be performed in an electronic learning environment, if you want your students to take the task from their computer. In our example, the EVS course, groups of students had to write a report on sustainable development and

enlargement of the European Union. This report had to contain useful advice for the European Committee concerning policy of water management, agriculture, or energy.

After the assessment task is determined, the elements of the task that determine the quality of the student's performance are defined. Sometimes, these can be found in so-called job profiles. Although these resources might prove to be very useful, they often include *lists of criteria* that could include too many skills or concepts or might not fit exactly. Most of the time, teachers develop their own criteria. A teacher has to analyse skills or products to identify performance criteria upon which to judge achievement. This is not easy. It is useful to use expert products or good examples to define the appropriate criteria. Communicating information about performance criteria provides a basis for the improvement of that performance. Quellmalz (1991) offers a set of specific guidelines for the development of quality performance criteria. Criteria should be significant, specifying important performance components, and represent standards that would apply naturally to determine the quality of performance when it typically occurs. The criteria must be clear and understandable for all persons involved. In e-learning environments, the teacher can determine these criteria in negotiation with students, in for example, discussion groups. The first column of Table II (adapted from Prins et al., 2005) shows the set of criteria used in the EVS course for formative as well as summative assessment. The criteria list was developed by one of the EVS teachers together with the first two authors of this article according to the abovementioned guidelines. Criteria concerned the product and the group process. Students in the EVS course were given the opportunity to negotiate about these criteria and to adjust them.

The final step is the creation of *performance scoring rubrics* that are used for formative and summative purposes. Contrary to more traditional forms of testing, performance assessments in which the students often are confronted with ill-defined problems, do not provide clear-cut right or wrong answers. The performance is evaluated in a way that allows informative scoring on multiple criteria. In a performance scoring rubric, the different levels of proficiency for each criterion should be defined. Using the information of the assessment form, feedback is given on a student's performance either in the form of a narrative report or a grade. A criterion-referenced qualitative approach is desirable, whereby the assessment will be carried out against the previously specified performance criteria. An analytic or holistic judgement then is given on the basis of the standard that the student has achieved on each of the criteria. Analytic or holistic rubrics that specify clear benchmarks of performance at various levels of proficiency also serve the purpose of guiding students as they perform the task (Birenbaum, 2003). Nystrand, Cohen, and Downing (1993) and

TABLE II
The Performance Scoring Rubric in EVS, Concerning Content and Process

Criteria	Above standard	At standard	Below standard	Attribute points earned
Content-related criteria				
	10–9	8–6	5–0	
Sustainable development is made operational.	Students give a definition used in their report and give practical tools to measure their solutions on these points.	Students give a definition used in their report but do not give practical tools to measure their solutions on these points or vice versa.	Students do not give a definition used in their report and do not give practical tools to measure their solutions on these points.	/10
	5	4–3	2–0	
The different aspects of SD are used in coherence.	The ecological, social and economic aspects of Sustainable Development are used in coherence and balance. Arguments are given for priority.	Not all aspect of sustainable development are used, but the ones that are used are in balance and coherent.	The different aspects of sustainable development are not used coherently or balanced.	/5
	10–9	8–6	5–0	
Consistency of the content, awareness of lacuna	In the different chapters, the same definitions are used, there is no overlap between the different chapters and the content of one chapter is not contradicting the content of another chapter. Insight is given into lacuna in knowledge.	In the different chapters, the same definitions are used, there is no overlap between the different chapters and the content of one chapter is not contradicting the content of another chapter. Lacunas in knowledge are disguised.	Different definitions are used. Chapters are contradicting the content of another chapter. Lacunas in knowledge are disguised.	/5

(*Continued on next page*)

TABLE II
(Continued)

Criteria	Above standard	At standard	Below standard	Attribute points earned
Content-related criteria				
	5	4–3	2–0	
Integration of disciplinary contributions	The different disciplines are integrated in each chapter and not only at the end.	The different disciplines are only integrated at the end of the report.	The different disciplines are not integrated.	/5
	10–9	8–6	5–0	
Relation between problem definition, analysis and solution	Scientific quality of report and conclusions and recommendations come from the chapters in the report. Questions asked in the beginning are answered.	Scientific quality of report and conclusions and recommendations come from the chapters in the report. Not all questions asked in the beginning are answered.	Low scientific quality of report and conclusions and recommendations do not come from the chapters in the report. Not all questions asked in the beginning are answered.	/10
	5	4–3	2–0	
Application of results	A target group is distinguished, is involved in the process and is ready to work on the applicability of the results.	A target group is distinguished, is involved in the process. It is not clear in which way the target group will work further with the results.	A target group is distinguished but is not involved in the process.	/5
	5	4–3	2–0	
Quality of language used	The style of the different chapters is the same and the English used is of good quality.	The style of the different chapters is different. The English used is of good quality.	The style of the different chapters is different. The English used is of bad quality.	/5

(*Continued on next page*)

TABLE II
(Continued)

Criteria	Above standard	At standard	Below standard	Attribute points earned
Content-related criteria				
	5	4–3	2–0	
Creativity	The different knowledge is linked in a creative way. The recommendations are provocative and sharp.	The different knowledge is linked in a creative way.	No new insight is given because knowledge of different disciplines and sources are left apart.	/5
	10–9	8–6	5–0	
Summary (separately!)	A 2–3 page summary is added, with: background research, recommendations, target group, possible implementation route. The summary is sharp, and provocative.	Summary is lacking one of the four points mentioned or leaves room for interpretation.	Summary is lacking two or more of the four points mentioned and leaves room for interpretation. Or no summary is added at all.	/10
	5	4–3	2–0	
Planning research	Not all work was done at the end. The spread was reasonable.	Most of the work was done at the end.	One or more of the deadlines have not been met.	/5
	5	4–3	2–0	
Planning individual tasks	The task division was clear and every student had a reasonable task.	The task division was clear but not every student had a reasonable task.	The task division was not clear.	/5
	5	4–3	2–0	
Cooperation within the group	Decisions were made together and every group member has a vote in the group.	Decisions were made but not every group member has a vote in it.	The group did work as a group. So no common decisions were made.	/5

(*Continued on next page*)

TABLE II
(Continued)

Criteria	Above standard	At standard	Below standard	Attribute points earned
Content-related criteria				
	5	4–3	2–0	
Cooperation via the internet	The internet is used for cooperation so that decisions made are traceable.	The internet is not always used for cooperation. Not all decisions made are traceable.	The group did not use internet for decisions.	/5
	5	4–3	2–0	
Participation	Each group member participated equally. Visits to internet, input into the report are equal.	Not every group member participated equally. Visits to internet, input into the report differs, but stays within reasonable variety.	Not every group member participated equally. Visits to internet, input into the report differs widely and caused problems in the group.	/5
	5	4–3	2–0	
Incorporating comments	The project team dealt with the comments, given by other teams, by the staff and the target group in a way that is recognisable and that fits with the rest of the report.	The project team dealt with the comments, given by other teams and by the staff in a way that is recognisable and that fits with the rest of the report. Comments from the target group are left out of the report.	The project team did not deal with the comments.	/5

Pitts, Coles and Thomas (2001) investigated whether it is preferable to have a holistic approach in performance assessment. When competencies are assessed through a task that requires the learners to integrate them, 'holistic' or 'integrated' assessment is required. This form of assessment relates to the whole unit or grouping of units, and requires observation of performance, questioning and, in some cases, review of documentation or other forms of evidence. The performance scoring rubrics used in the EVS course are presented in Table II. One EVS teacher and the first two authors of this article designed two performance scoring rubrics, one for the product and one for the collaboration process. These rubrics were evaluated by the five other teachers and the co-ordinator of the EVS course. The student groups used the rubric concerning the product for the formative assessment of the first draft of a report of a fellow group, whereas the teachers used both rubrics for the summative assessment of the revised final draft of the group report and the collaboration process. The scoring rubric for the product counted for 70% of the end mark (9 criteria, see Table II) whereas the scoring rubric for the group process counted for 30%.

2. Assuring Quality In Performance Assessment

In CBL, it is important that a number of performance assessments are organised to gather reliable and valid information about a learner's competency development. The standard approaches to reliability and validity are derived from a psychometric approach, which is based upon the notion of reaching an 'ideal' grade (Johnston, 2004). In competency-oriented learning contexts, Dierick, van de Watering and Muijtjens (2002) indicate a shift from psychometric to edumetric criteria for the quality of assessment scores. There is more attention for criteria like accuracy of the scores, the cognitive complexity of the performance task, the authenticity of the performance task, the transparency of the assessment procedure, and the fairness in assessment. Each performance assessment, however, has a weak link in the quality chain that links the performance of the learner and the conclusion about the competency in a particular context (Crooks, Kane, & Cohen, 1996). To tackle this problem, three important criteria are specifically important to cover in the design of performance assessments: accuracy, generalisibility, and extrapolation.

A performance assessment is *accurate* when the score comes close to the true score of the learner. The true score is a theoretical concept defined as the mean score of an infinite number of measurements by means of equivalent assessments, assuming that there are no changes in the person or any other effects. The assessment of competencies also implies more than one observed performance. The learner has to perform similar types

of tasks in a variety of situations under the same conditions. Studies generally conclude that the *generalisibility* to performances in similar tasks is limited (Linn et al., 1991). The main reason for this finding is that the assessments in current learning environments are a poor reflection of all possible tasks that, in fact, could be presented to the learner (probably due to lack of time and money). It is therefore recommended that a variety of performance assessments that represent a certain level of authenticity be defined (Gulikers, Bastiaens, & Martens, 2005). *Extrapolation* implies that the attained score reflects the performance level that the learner would achieve in a real working situation. Sometimes this is no problem, because the assessment task does not deviate from the task in the real situation. But often it is a problem. For example, when the performance task is too expensive (launch of a Patriot rocket), too dangerous (defusing a bomb), or when the situation is unlikely to occur in real life (the arrest of an armed criminal in a shopping centre). In most assessments, the level of realism (i.e. 'fidelity') is reduced. The more the fidelity is reduced, the more difficult it is to prove that the attained score is a realistic reflection of the authentic performance in the working field.

The three quality criteria place heavy demands on the design and organisation of performance assessments, but they are also problematic in the sense that optimising one criterion leads to an impairment of another criterion. Therefore, it is important to choose a design approach for learning and assessment that warrants for all the quality aspects (Straetmans & Sanders, 2001). A model that provides guidelines to design competency-based education, that focuses on the design of whole tasks, in which instruction, learning, and assessment can be aligned, is the Four Component Instructional Design model (4C/ID model), developed by van Merriënboer (van Merriënboer, 1997; van Merriënboer, Jelsma, & Paas, 1992). We try to illustrate its usability when it comes to the design of good and valid assessment.

3. Instructional Design For Competency-Based Performance Assessment

In the 4C/ID model, competencies are defined as complex skills, consisting of integrated sets of constituent skills with their underlying knowledge structures and attitudes (van Merriënboer, 1997). Examples of complex skills are delivering training (consultant), designing a house (architect), or supervising a public domain (police officer). The basic components of the model are presented in Figure 1.

The *tasks* (first component) are the backbone of every educational program aimed at the acquisition of competencies (see Figure 1, which represents the tasks as circles). The tasks are typically performed in a real

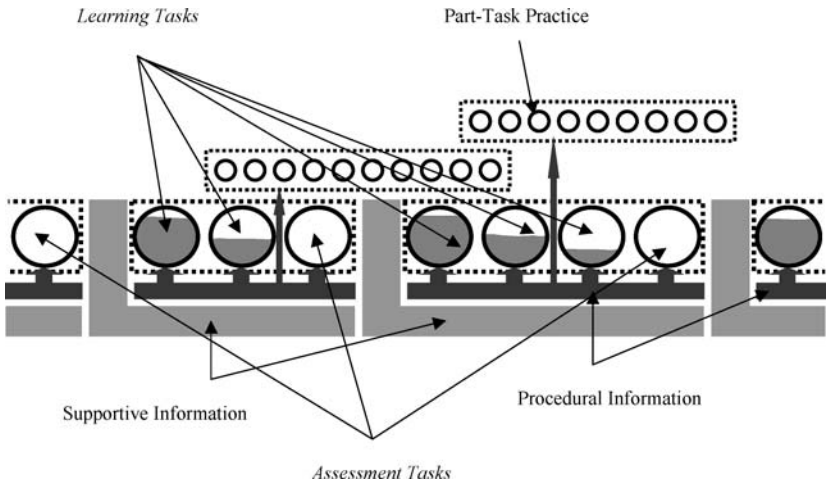

Figure 1. The four components in the 4C/ID model.

or simulated task environment and provide 'whole-task practice': ideally, they confront the learners with all constituent skills that make up the whole competency. The tasks in the EVS course are good examples of an authentic task in a simulated task environment that gives the students the opportunity of whole-task practice. Students had to conduct research in a multidisciplinary team with team members of different nationalities, as in situations that they could encounter after their study.

Learners will typically start their study on relatively simple tasks and progress towards more complex tasks. Complexity is affected by the amount of skills involved, the amount of interactions between skills, and the amount of knowledge necessary to perform these skills. Task classes are used to define simple-to-complex categories of tasks and to steer the process of selection and development of suitable tasks (see the dotted lines around the circles in Figure 1). Tasks within a particular task class are equivalent in the sense that the tasks can be performed on the basis of the same body of knowledge. The basic idea is to use a whole-task approach in which the first task class refers to the simplest version of whole tasks that professionals encounter in the real world. For increasingly more complex task classes, the assumptions that simplify task performance are relaxed. The final task class represents all tasks, including the most complex ones that professionals encounter in the real world. The task of the EVS course was rather complex, considering the skills and knowledge that are needed to conduct the research and write the report. Obviously, the EVS course is not appropriate for students at the start of their curriculum.

Once the task classes are defined, the tasks can be selected and/or developed for each class. The cases that are selected for each task class form the

basis for the to-be-developed tasks. For each task class, enough cases are needed to ensure that learners receive enough practice to reach mastery. It should be noted that the cases or tasks within the same task class are not further ordered from simple to complex; they are considered to be equivalent in terms of difficulty. A high variability of the tasks within the same task class is of utmost importance to facilitate the development of generalised cognitive schemata and reach transfer of learning (e.g. Gick & Holyoak, 1983; Paas & van Merriënboer, 1994). In fact, the EVS course contained four tasks, that is, four cases of equal complexity (the cases concerned agricultural policy, integrated water management, energy technology, and spatial planning and policy). For practical reasons, the student groups in the EVS course wrote one group report based on one case, although it might have been better for transfer and development of schemata if students had done two or more cases.

While there is no increasing difficulty for the tasks within one task class, they do differ with regard to the amount of support provided to learners. Much support is given for tasks early in each task class, which therefore are labelled as learning tasks, and this support diminishes until no support is given for the final learning task in a task class (see the filling of the circles in Figure 1). The last unguided and unsupported tasks in a task class (i.e. the empty circles) are therefore suitable performance assessment tasks (see also Figure 1). This task is designed according to guidelines of Stiggins (1987) as outlined in a previous section. The assessment task focuses on the ability to use combinations of acquired skills, knowledge, and attitudes and therefore fits in well with the theory of competency-based learning environments (Linn et al., 1991). In the EVS course, support like worked-out examples or process worksheets was limited, which makes this task suitable for performance assessment. The idea is that most of the learning should be done during earlier courses in the curriculum.

Obviously, learners need information in order to work fruitfully on learning tasks and to learn genuinely from those tasks. This *supportive information* (second component) provides the bridge between what learners already know and what they need to know to work on the learning tasks. It is the information that teachers typically call 'the theory' and which is often presented in study books and lectures. Because the same body of general knowledge underlies all learning tasks in the same task class, and because it is not known beforehand which knowledge precisely is needed to successfully perform a particular learning task, supportive information is not coupled to individual learning tasks but to task classes (see the 'supportive information' in Figure 1). In the EVS course, supportive information was provided by documents in the VLE and by domain experts connected to the EVS course.

Whereas supportive information pertains to the non-recurrent aspects of a complex skill, *procedural information* (third component) pertains to the recurrent aspects, that is, constituent skills of a competency that should be performed after the training in a highly similar way over different problem situations. Procedural information provides learners with the step-by-step knowledge that they need to know in order to perform the recurrent skills. They can be in the form of, for example, directions that teachers or tutors typically give to their learners during practice, acting as an 'assistant looking over your shoulder' (ALOYS), information displays, demonstrations or feedback. Because procedural information is identical for many tasks, which all require the same recurrent constituent skills, it is typically provided during the first learning task for which the skill is relevant (see 'procedural information' in Figure 1). In the EVS course, procedural information was of minor importance because this type of information should have been provided during courses earlier in the curriculum.

Finally, if a very high level of automaticity of particular recurrent aspects is required, the learning tasks could provide insufficient repetition to provide the necessary amount of practice to reach this level. Only then, it is necessary to include additional *part-task practice* (fourth component) for those selected recurrent aspects in the training program (see 'part-task practice' in Figure 1). This was not the case in the EVS course.

When learners work on an assessment task in a particular task class, they have to show their progress on both the recurrent aspects of performance, which are routines that are consistent from problem to problem situation, and the non-recurrent aspects of performance, which involve problem solving or reasoning and vary over situations.

Figure 2 depicts how performance assessment can be intertwined with the four components. In general, different assessment methods are appropriate for each of the components (see van Merriënboer, 1997). The performance scoring rubric can be a valuable tool to provide formative feedback to students. For summative assessment, the performance scoring rubric helps the teacher to reach a balanced grading and final decision. It is up to the teacher or designer to decide whether each assessment task (see Figure 1) is used for summative assessment or, for example, only the assessment task at the end of the most complex task class.

From a theoretical viewpoint, assessment of whole-task performance is the only form of assessment that is unconditionally required in integrated e-learning or any other educational setting for complex learning. The 4C/ID model states that students cannot satisfactorily perform such whole assessment tasks if they do not possess the mental models and cognitive strategies (i.e. theoretical knowledge) that help them to perform the non-recurrent aspects of the task and the procedures or rules that govern the performance of the recurrent aspects of the task. Nonetheless, additional assessment of

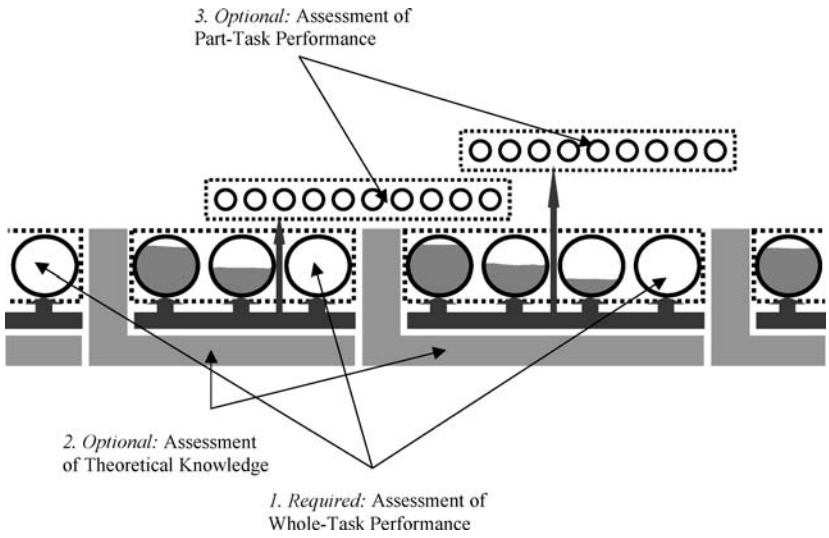

Figure 2. A framework for designing performance assessment in integrated e-learning.

theoretical knowledge can be applied for a number of reasons. First of all, it could help in diagnosing students' conceptual problems or misconceptions and yield the necessary information to give them formative feedback for overcoming these problems. And, furthermore, it might be used to corroborate the findings from the assessment of whole-task performance, making the whole assessment more reliable.

Like the assessment of theoretical knowledge, the assessment of part-task performance on single recurrent skills also can be considered as an additional element in the whole assessment system. Preferably, the same tools that are used for part-task practice are also used for the assessment of the recurrent skill under consideration. Most drill-and-practice computer programs (e.g. for using grammatical rules in second-language learning; applying safety procedures in industry, or operating particular software packages in business) indeed assess students on their accuracy and speed, and use this information to diagnose errors, to indicate to students that there is an error, and to provide hints that could help students to get back on the right track. Thus, for collecting evidence concerning theoretical knowledge or recurrent skills, traditional tests could be helpful. In the EVS course, the required assessment of whole-task performance was executed using the performance scoring rubrics. The teachers of the EVS course decided not to assess part-task practice and theoretical knowledge.

Concluding, we argue that one should always focus on performance assessment of whole tasks. The definition of those assessment tasks early in the design process might also be helpful for the development of appropriate learning tasks that guide students towards the assessment task(s) at the

end of a task class. Furthermore, one might consider including additional assessments for theoretical knowledge and for recurrent skills that have been separately practised.

4. Discussion

This article focuses on the integration of instruction and performance assessments in e-learning. An approach for designing competency-based instruction and performance assessments with a specific focus on e-learning contexts is presented, which implies an adjusted view of instructional design, learning and assessment. It is argued that one should always focus on performance assessment of whole tasks. The definition of those assessment tasks early in the design process is helpful for the development of appropriate learning tasks that guide students towards the assessment task(s) at the end of a study unit.

However, some possible pitfalls are conceivable and could occur when working with this approach. This discussion briefly addresses these issues. Specific conditions for successful implementation of assessment in e-learning have to be met. The core of this applies to many e-learning situations: students often don't do the things that designers or teachers expect them to (Jochems, van Merriënboer, & Koper, 2004; Lockwood, 1995; Martens, 1998; Mudrack & Farrell, 1995; Strijbos, Martens, & Jochems, 2004), and this is of great concern. No matter how high the quality of an assessment procedure is, there is no assurance that students learn in the intended way. There is a distinction between 'what is meant to happen', that is, the curriculum stated officially by the educational system or institution, and what teachers and learners actually do and experience 'on the ground', which is a kind of de facto curriculum. Snyder (1973) labels this the 'hidden curriculum'. In a laboratory, researchers can ask students to read texts but, in 'real life', students have their own hidden curriculum, "adopting ploys and strategies to survive in the system" (Lockwood, 1995, p. 197). The solution to this problem could be the improvement of students' intrinsic motivation. Ryan and Deci (2000) put forward a model that indicates that certain aspects of the social environment and task environment influence student motivation. For example, intrinsic motivation can be influenced by relatedness or trust in peers and by stimulating students to work on assessment tasks with authenticity that are strongly related to the learning tasks (Furrer & Skinner, 2003; Ryan & Deci, 2000).

Even when teachers and educational developers manage to solve motivational problems, there are more possible problems that need to be solved. A risk factor of integrated performance assessment in an e-learning setting is that the design of these assessments puts heavy demands on teachers and

developers (e.g. Beijaard, Verloop, Wubbels, & Feiman-Nemser, 2000; Sluijsmans, Moerkerke, Dochy, & van Merriënboer, 2001). Introducing e-learning and performance assessment is difficult, requires new teacher roles, requires them to collaborate with many stakeholders and can be time consuming. Moreover, students need to be convinced of the usefulness of competency-learning contexts. If students are not convinced of the usefulness of performance assessment and are not sufficiently intrinsically motivated, it is unlikely that the performance assessment will become a success. The perception of the learning environment might play a mediating role in the interplay between students' intended study strategies, their perceptions of the assessment demands and their actual learning strategies (Nijhuis, Segers, & Gijselaers, 2005).

Still, we are convinced that, in spite of these possible risks and problems, the benefits of competency-based performance assessment are too great to discourage the implementation of performance assessment in e-learning. In general, the implementation of performance assessments in e-learning holds a number of important advantages (Surgue, 2002). The advantages are related to the integration of assessment and instruction, the possibilities for adequate feedback, the involvement of students, and the authenticity of performance assessments. When looking at the implementation in e-learning, Baker and Mayer (1999) state that computers can have three-fold value in web-based performance assessment. First, computers have the ability to capture process differences. It is possible to trace back indicators that provide information about which thinking processes contributed to a particular performance. Second, computers can make complex processes visible. And, third, online scoring and feedback can be provided based on fixed moments or on a student model. A student model is based on the logging of actions that students conduct during their learning.

Especially in e-learning, learners can play a valuable part in performance assessment by means of peer assessment. Peer assessment implies that students evaluate the performances of peers and provide constructive feedback (Sluijsmans, 2002). At least three arguments support the implementation of peer assessment in e-learning. First, integrating peer assessment supports students with their development into competent professionals who continuously reflect on their behaviour and their learning. There seem to be several ways in which students can be involved in assessment. First, students can have a role in the choice of performance assessment tasks and in discussing assessment criteria (Mehrens, Popham, & Ryan, 1998). Second, it is substantiated that peer assessment promotes integration of assessment and instruction, making the student an active person who shares responsibility, reflects, collaborates and conducts a continuous dialogue with the teacher. Third, peer assessment can decrease the workload of teachers. In the EVS course, peer assessment was integrated in the tasks by letting

student assess the quality of the first draft of the report of a fellow group. This led, for instance, to better final drafts of the group reports and, thus, to a decrease of teacher workload.

The advantages of e-learning that are often mentioned, such as ease of distribution, timeliness, immediate feedback, variety of delivery modes, tracking, long-term cost savings and convenience, are mostly true for item-based tests, but less applicable for competency-based performance assessments, where the benefits are predominantly based on educational grounds. Item-based tests can be valuable for assessment of part-task practice, but are not useful for whole-task assessment.

In line with Birenbaum (2003), we conclude that much more research is required to better understand the nature of competency-based performance assessment in e-learning and the impact on learning. For instance, further study is needed into the negotiating process whereby assessment criteria are set, the process by which students come to internalise the standards of good performance and the impact of motivational processes in general. Also, the role of teachers and the complex process of curriculum redesign need to be addressed (Sluijsmans, 2002). In our opinion, performance assessment is a crucial factor in educational innovation. When students are really motivated to perform, study, learn and collaborate in a new way, and if learning goals and learning processes are much more in line, educational problems such as lack of motivation, early drop-out, and test behaviour might be decreased.

REFERENCES

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

Baker, E. L., & Mayer, R. E. (1999). Computer-based assessment of problem solving. *Computers in Human Behavior*, *15*, 269–282.

Bastiaens, Th., & Martens, R. (2000). Conditions for web-based learning with real events. In B. Abbey (Ed.), *Instructional and cognitive impacts of web-based education* (pp. 1–32). Hershey/London: Idea Group Publishing.

Beijaard, D., Verloop, N., Wubbels, Th., & Feiman-Nemser, S. (2000). The professional development of teachers. In R. J. Simons, J. van der Linden, & T. Duffy (Eds.), *New learning* (pp. 261–274). Dordrecht, The Netherlands: Kluwer.

Bennett, R. E. (1999). Using new technology to improve assessment. *Educational Measurement: Issues and Practice*, *18*, 5–12.

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

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.

Crooks, T. J., Kane, M. T., & Cohen, A. S. (1996). Threats to the valid use of assessments. *Assessment in Education*, *3*, 265–285.

Dierick, S., van de Watering, G., & Muijtjens, A. (2002). De actuele kwaliteit van assessment: Ontwikkelingen in de edumetrie [Current quality in assessment: Developments in edumetrics]. In F. Dochy, L. Heylen, & H. van de Mosselaer (Eds.), *Assessment in onderwijs: Nieuwe toetsvormen en examinering in het studentgericht onderwijs en competentiegericht onderwijs* [Assessment in education: New modes of assessment in student-centred and competency-based education] (pp. 91–122). Utrecht, The Netherlands: Lemma.

Furrer, C., & Skinner, E. (2003). Sense of relatedness as a factor in children's academic engagement and performance. *Journal of Educational Psychology*, *95*, 148–162.

Garris, G., Ahlers, R., & Driskell, J. (2002). Games, motivation, and learning. S*imulation & Gaming*, *33*, 441–467.

Gick, M., & Holyoak, K. J. (1983). Schema induction and analogical transfer. *Cognitive Psychology*, *15*, 1–38.

Gulikers, J., Bastiaens, Th., & Martens, R. (2005). The surplus value of an authentic learning environment. *Computers in Human Behavior*, *21*, 509–521.

Herman, J. L., Aschbacher, P. R., & Winters, L. (1992). *A practical guide to alternative assessment*. Alexandria, VA: Association for Supervision and Curriculum Development.

Jochems, W., van Merriënboer, J., & Koper, R. (2004). *Integrated e-learning: Implications for pedagogy, technology & organization*. London: RoutledgeFalmer.

Johnston, B. (2004). Summative assessment of portfolios: An examination of different approaches to agreement over outcomes. *Studies in Higher Education*, *29*, 395–412.

Linn, R. L., Baker, E. L., & Dunbar, S. B. (1991). Complex, performance-based assessment: Expectations and validation criteria. *Educational Researcher*, *20*(8), 15–21.

Lockwood, F. (1995). Students' perception of, and response to, formative and summative assessment material. In F. Lockwood (Ed.), *Open and distance learning today* (pp. 197–207). London: Routledge.

Martens, R. L. (1998). *The use and effects of embedded support devices in independent learning* (PhD thesis). Utrecht, The Netherlands: Lemma BV.

Mayer, R. E. (2002). A taxonomy for computer-based assessment of problem solving. *Computers in Human Behavior*, *18*, 623–632.

Mehrens, W. A., Popham, W. J., & Ryan, J. M. (1998). How to prepare students for performance assessments. *Educational Measurement: Issues and Practice*, *17*, 18–22.

Mudrack, P. E., & Farrell, G. M. (1995). An examination of functional role behaviour and its consequences for individuals in group settings. *Small Group Research*, *26*, 542–571.

Nijhuis, J. F. H., Segers, M. S. R., & Gijselaers, W. H. (2005). Influence of redesigning a learning environment on student perceptions and learning strategies. *Learning Environments Research*, *8*, 67–93.

Nystrand, M., Cohen, A. S., & Dowling, N. M. (1993). Addressing reliability problems in the portfolio assessment of college writing. *Educational Assessment*, *1*, 53–70.

O'Neil, H. F. (2002). Perspectives on computer-based performance assessment of problem solving. *Computers in Human Behavior*, *15*, 255–268.

Paas, F. G. W. C., & van Merriënboer, J. J. G. (1994). Variability of worked examples and transfer of geometrical problem solving skills: A cognitive load approach. *Journal of Educational Psychology*, *86*, 122–133.

Pitts, J., Coles, C., & Thomas, P. (2001). Enhancing reliability in portfolio assessment: 'Shaping' the portfolio. *Medical Teacher*, *23*, 351–365.

Prins, F. J., Sluijsmans, D. M. A., Kirschner, P. A., & Strijbos, J. W. (2005). Formative peer assessment in a CSCL environment: A case study. *Assessment and Evaluation in Higher Education*, *30*, 417–444.

Quellmalz, E. (1991). Developing criteria for performance assessments: The missing link. *Applied Measurement in Education*, *4*, 319–332.

Reeves, T. C. (2000). Alternative assessment approaches for online learning environments in higher education. *Journal of Educational Computing Research*, *23*, 101–111.

Ryan, R. M., & Deci, E. L. (2000). Self-determination theory and the facilitation of intrinsic motivation, social development, and well being. *American Psychologist*, *55*, 68–78.

Segers, M., & Dierick, S. (2001). Quality Standards for new modes of assessment: An exploratory study of the consequential validity of the OverAll test. *European Journal of Psychology of Education*, *16*, 569–588.

Shavelson, R. J. (1991). Performance assessment in science. *Applied Measurement in Education*, *4*, 347–362.

Sluijsmans, D. M. A. (2002). *Student involvement in assessment: The training of peer assessment skills*. Unpublished doctoral dissertation, Open University of The Netherlands, Heerlen, The Netherlands.

Sluijsmans, D., Moerkerke, G., Dochy, F., & van Merriënboer, J. (2001). Peer assessment in problem based learning. *Studies in Educational Evaluation*, *27*, 153–173.

Snyder, B. (1973). *The hidden curriculum*. Cambridge, MA: The MIT Press.

Straetmans, G. J. J. M., & Sanders, P. F. (2001). *Beoordelen van competenties van docenten* [Assessment of competencies of teachers]. Den Haag, The Netherlands: Programmamanagement ESP/HBO-raad.

Stiggins, R. (1987). Design and development of performance assessment. *Educational Measurement: Issues and Practice*, *6*, 33–42.

Stoof, A., Martens, R., van Merriënboer, J., & Bastiaens, Th. (2002). The boundary approach of competence: A constructivist aid for understanding and using the concept of competence. *Human Resource Development Review*, *1*, 345–365.

Strijbos, J. W., Kirschner, P. A., & Martens, R. L. (Eds.). (2004). *What we know about CSCL in higher education*. Dordrecht, The Netherlands: Kluwer.

Strijbos J. W., Martens, R. L., & Jochems, W. M. G. (2004). Designing for interaction: Six steps to designing computer supported group based learning. *Computers & Education*, *42*, 403–424.

Surgue, B. (2002). Performance-based instructional design for e-learning. *Performance Improvement*, *41*, 45–50.

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 training complex cognitive skills. *Educational Technology, Research and Development*, *40*, 23–43.

Van Merriënboer, J. J. G., & Martens, R. L. (2002). Computer-based tools for instructional design. *Educational Technology, Research and Development*, *50*, 5–9.

Wiggins, G. (1989). *A true test: Toward a more authentic and equitable assessment*. Phi Delta Kappan, 70, 703–713.

Zhang, J., Khan, B. H., Gibbons, A. S., & Ni, Y. (2001). Review of web-based assessment tools. In B. H. Khan (Ed.), *Web-based training* (pp. 287–295). Englewood Cliffs, NJ: Educational Technology Publications.

DOMINIQUE M. A. SLUIJSMANS AND FRANS J. PRINS
Open University of The Netherlands
Educational Technology Expertise Centre,
P.O. Box 2960
6401 DL Heerlen, The Netherlands
E-mails: dominique.sluijsmans@ou.nl; frans.prins@ou.nl

ROB L. MARTENS
Centre for the Study of Education and Instruction
Faculty of Social and Behavioural Sciences
Leiden University
P.O. Box 9555
2300 RB Leiden, The Netherlands
E-mail: rmartens@fsw.leidenuniv.nl

(Correspondence to: Dominique M. A. Sluijsmans.
E-mail: dominique.sluijsmans@ou.nl)