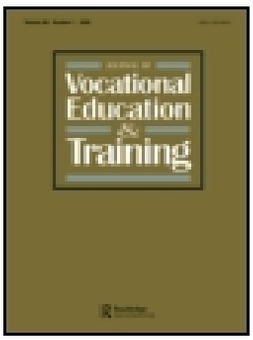

Routledge
Taylor & Francis Group

Journal of Vocational Education and Training

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

Integrating authentic assessment with competence-based learning in vocational education: the Protocol Portfolio Scoring

Dominique M.A. Sluijsmans , Gerard J.J.M. Straetmans & Jeroen J.G. van Merriënboer

To cite this article: Dominique M.A. Sluijsmans , Gerard J.J.M. Straetmans & Jeroen J.G. van Merriënboer (2008) Integrating authentic assessment with competence-based learning in vocational education: the Protocol Portfolio Scoring, Journal of Vocational Education and Training, 60:2, 159-172, DOI: 10.1080/13636820802042438

To link to this article: https://doi.org/10.1080/13636820802042438

Published online: 30 Apr 2008.

Journal of Vocational Education and Training
Vol. 60, No. 2, June 2008, 159–172

Routledge
Taylor & Francis Group

Integrating authentic assessment with competence-based learning in vocational education: the Protocol Portfolio Scoring

Dominique M.A. Sluijsmans[a]*, Gerard J.J.M. Straetmans[b] and Jeroen J.G. van Merriënboer[a]

[a]*Open University of the Netherlands;* [b]*CITO, Arnhem, The Netherlands*

(*Received 27 July 2007; final version received 7 February 2008*)

This article describes how competence-based learning (CBL) can be organised in vocational education by integrating elements from a holistic instructional design model with recent ideas on assessment. A curriculum based on this model is pre-eminently suitable for an assessment approach emphasising that proof of competence is gathered by having learners perform authentic tasks under changing assessment conditions at regular intervals. The results are stored in a so-called electronic assessment portfolio. The portfolio is constructed according to the Protocol Portfolio Scoring (PPS). The value of PPS for flexible, demand-driven vocational education is discussed.

Keywords: authentic assessment; competence-based learning; portfolio assessment; vocational education

Introduction

In response to the labour-market demand for motivated workers, institutions for secondary vocational education adopted competence-based learning (CBL) and redesigned their learning environments accordingly (Tillema, Kessels, and Meijers 2000). Characteristic of CBL are real-life complex problems that require students to actively engage in their learning. These real-life problems help learners to integrate the competences necessary for effective task performance in work settings (Stoof et al. 2002; Gulikers, Bastiaens, and Kirschner 2004). This competence-based learning (CBL) may sound like the answer to the changing requirements in the labour market, but often the assessment methods are not adapted to this philosophy. Yet, it is well known that assessment has a strong influence on the quality of student learning. 'Do we need to know this for the exam?' is still one of most frequently asked questions in the classroom (Sluijsmans 2002), illustrating the 'washback effect': assessment strongly influences the study behaviour of learners' assessment and overrides practically every other aspect of curriculum design (Frederiksen 1984; Alderson and Wall 1993).

Despite many efforts in educational innovation to establish CBL, there are still some flaws in the current assessment systems in CBL. First, we see that the assessment is often designed apart from the instructional activities. Since teachers are not educated in how to design assessments, they first design the instructional activities and then start to think about the assessment. As a consequence, the lack of alignment between the instruction and the assessment often leads to students being surprised when they are confronted with the assessment with comments like 'I expected something else' or 'The questions on the exam were totally different from the tasks we undertook in the course'. Furthermore we see that teachers in vocational education are led by a

*Corresponding author. Email: dominique.sluijsmans@ou.nl

ISSN 1363-6820 print/ISSN 1747-5090 online

DOI: 10.1080/13636820802042438
http://www.informaworld.com

norm-based assessment approach, where a student's learning outcome is compared with a mean score of their peer group. These norm-based exams, however, assume a uniformity of learners, and reinforce learners' test behaviour and teachers' teaching to the test. In fact, the assessment does not contribute to learning in the long term.

Unfortunately, to date the instruction and assessment in vocational education is mainly externally controlled by the teacher or the school system. What, when and how things should be learned is prescribed. For assessment, this situation encourages students to adopt a passive attitude, only absorbing the assessment results without questioning the function of the assessment or the 'why' behind the teacher's provision of empty, one-dimensional grades. Although research has proved that grading is the least powerful mode of feedback, it is still common assessment practice. Rich, formative modes of assessment that foster learners' further learning (see Black and Wiliam 1998, for an overview) are not naturally interwoven in daily educational practice. Worse still, the assessments are often just one-shot measurements, although there is evidence that for a reliable and valid judgement regarding a student's performance, multiple assessments are necessary on different levels of complexity.

For CBL to succeed, it is necessary to use different kinds of assessments in which learners are not tested solely on their remembering of knowledge, but more on their ability to interpret, analyse and evaluate problems and explain their arguments. These assessments are always based on criteria and are thus 'criterion referenced', in comparison with norm-referenced assessment in which individual performance is compared with that of a larger group. Thus, CBL is more likely to succeed if learning, instruction and assessment are constructively aligned (Arter 1996; Biggs 1996; Dochy and McDowell 1997). Pursuing the theory of constructive alignment, we argue that assessment should be regarded from an instructional design perspective that is based on the acquisition of competences (Birenbaum 2003). For this, we regard competence-based assessment as authentic assessment, defined by Gulikers as 'an assessment requiring learners to demonstrate the same [kind of] competencies … that they need to apply in the criterion situation in professional life' (Gulikers, Bastiaens, and Kirschner 2004, 5). Authentic assessment is also referred to as alternative assessment (as an alternative to 'traditional' forms of assessment; Baartman, Bastiaens, and Kirschner 2006) or performance assessment, because learners are asked to perform meaningful tasks. Some researchers distinguish performance assessment from authentic assessment by defining performance assessment as performance based but with no reference to the authentic nature of the task (e.g. Meyer 1992).

Unfortunately, no instructional design models are at hand that fully integrate CBL and modes of authentic assessment (van Merriënboer 1997; Straetmans et al. 2003). Therefore, we introduce an integrative framework for the design of CBL and authentic assessment, based on two well-grounded design approaches: the Four Component Instructional Design model (4C/ID model; van Merriënboer et al. 1992) and the Protocol Portfolio Scoring, a method for continuous monitoring of assessment results (Straetmans et al. 2003). The surplus value of the 4C/ID model and PPS for flexible CBL is addressed, with a specific focus on the self-regulation of learners.

Designing authentic assessment in CBL: the Four Component Instructional Design model

A model that provides guidelines to design CBL, in which instruction, learning and assessment are fully aligned, is the Four Component Instructional Design model (4C/ID model), originally developed by van Merriënboer, Jelsma, and Paas (1992). In the 4C/ID model competences 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 giving training (consultant), designing a house (architect), or supervising a public domain (police officer). The basic components of the model are presented in Figure 1. To illustrate the model

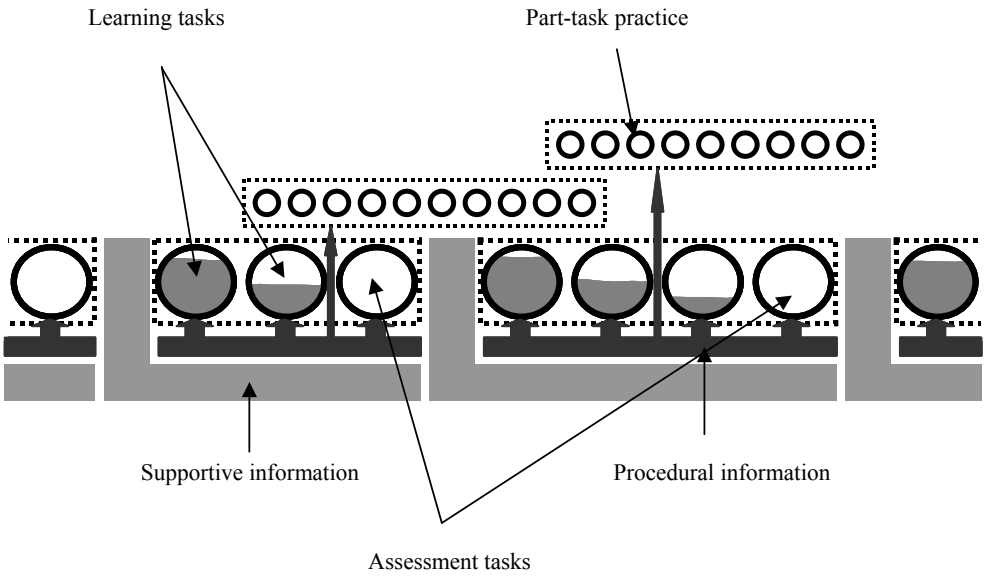

Figure 1. The four components in the 4C/ID-model.

and the organisation of authentic assessment, we pursue the example of the competence 'Supervise a public domain'.

The *tasks* (first component) are the backbone of every educational programme aimed at the acquisition of competences (see Figure 1, which represents the tasks as circles). The tasks are typically performed in a real or simulated task environment and provide 'whole-task practice': ideally, they confront the learners with all constituent skills that make up the whole competence. It is, however, clearly impossible to provide highly complex tasks right from the start because this would yield excessive cognitive load (or overload) for the learners, which impairs learning and performance (Sweller, van Merriënboer, and Paas 1998). Thus, learners will typically start their work on relatively simple tasks and progress towards more complex tasks. Complexity is affected by the amount of constituent skills involved, the amount of interactions between constituent skills, and the amount of knowledge necessary to perform the constituent skills. For the competence 'Supervise a public domain' some constituent skills are for example: (1) indicate deviant behaviour; (2) manage conflicts between two or more persons; and (3) use adequate conversation techniques. A competent police officer is able to integrate and coordinate these skills in a particular situation, and to apply these skills in new situations (i.e. transfer of learning).

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 where 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 the most complex situations that professionals encounter in the real world. A simple illustration of this simplifying assumptions approach for the competence 'Supervise a public domain' is presented in Table 1.

Once the task classes are defined, the tasks can be selected and/or developed for each class. For instance, one could ask an experienced police officer to come up with concrete cases in

Table 1. Examples of task factors for the complex skill 'supervise a public domain'.

Task factor	Task class 1 (simple tasks)	Task class 2	Task class 3 (complex tasks)
Number of people	20–80	80–200	> 300
Environment	Public domain with regular activity, e.g. a village square, no risks	Public domain with a special happening, e.g. a funfair at the village square, some risks	Public domain with lack of safety, e.g. a noisy place of entertainment, many risks
Incident	A minor violation, e.g. a cyclist in a pedestrians-only area	A theft, e.g. a bag snatcher or pick-pockets	A violent crime, e.g. a gun battle
Way of acting	Observing	Observing, giving advice for prevention, regulating	Observing, giving advice for prevention, regulating
Briefing	Not attending a briefing	Not attending a briefing	Attending a briefing
Research	Not starting a criminal investigation	Not starting a criminal investigation	Starting a criminal investigation

which he/she supervised complex areas, with a high risk violation, and how he/she defined advices for prevention (i.e. cases that fit within the last task class). The same is done for preceding, easier task classes. 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 and Holyoak 1983; Paas and van Merriënboer 1994).

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).

Obviously, learners need information in order to work fruitfully on learning tasks and to genuinely learn 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).

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 competence 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 they need to know in order to perform the recurrent skills. This can be in the form of, for example, directions teachers or tutors typically give to their learners during practice, acting as an 'assistant looking over your shoulder' (ALOYS), information displays, demonstrations or feedback. In the context of the police officer this could be the procedure for filling out a ticket for speeding. 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).

Finally, if a very high level of automaticity of particular recurrent aspects is required, the learning tasks may provide insufficient repetition to provide the necessary amount of practice to reach this. Only then is it necessary to include additional *part-task practice* (fourth component) for those selected recurrent aspects in the training programme (see 'part-task practice' in Figure 1). For the police domain this could be, for example, part-task training in shooting with a pistol.

The 'whole task' approach is essential for the assessment of competences. Because learners are directly confronted with realistic learning tasks at the start of the educational programme, information can be gathered that is useful in making judgements at the end of the educational programme concerning the level of competence. The last, unguided and unsupported tasks in task classes (i.e. the empty circles) are suitable as *assessment* tasks (see also Figure 1). These tasks are described in terms of a certain performance that is perceived as worthwhile and relevant to the learner and represents a whole task (Wiggins 1989). The assessment tasks, which can vary in level of authenticity (Gulikers et al. 2006), require learners to demonstrate the ability to use combinations of acquired skills, knowledge and attitudes and therefore fits CBL and the basic assumptions of the 4C/ID model (Linn, Baker, and Dunbar 1991). Results of the assessment tasks can be used to decide on a possible shift to a more difficult task class (is the learner ready for a next difficulty level?) or on the completion of the whole educational programme (has the learner acquired the specific competences?). The Protocol Portfolio Scoring allows for systematic and continuous monitoring of a learner's progress on multiple assessment tasks.

Protocol Portfolio Scoring: a new approach in portfolio assessment

Many educational programmes in vocational education settings have been using portfolios as a means for gathering evidence for the acquisition of competences. Research shows that portfolios are mostly used for formative purposes to foster reflection (e.g. Wade and Yarbrough 1996; Borko et al. 1997; Smith and Tillema 2000). Portfolios offer the chance to be a record of personal development. Psychometric data to support the use of portfolios as an assessment tool are, however, sparse and lacking in the scientific literature (Pitts et al. 2001). Although some researchers focus strongly on issues of reliability (e.g. Reckase 1995) and validity (e.g. Le Mahieu, Gitomer, and Fresh 1995), we see which others state that applying measures such as reliability and validity is not appropriate for portfolio assessment because this implies a reductionist view on assessment (Pitts et al. 2001). Reckase (1995) argues that because of low reliability of scores, portfolios should only be considered for formative purposes and not for summative purposes.

To underpin both validity issues of portfolios and a reliable *summative* use of portfolios, with regard to a more innovative instructional approach to assessment, Straetmans developed the Protocol Portfolio Scoring (further indicated as PPS; Straetmans et al. 2003). PPS provides guidelines how to collect pieces of evidence over time regarding a certain competence in a valid and reliable way. The four main goals of PPS are: (1) making sound decisions about the competence development of learners; (2) closing the gap between the artificial and redundant distinction between formative and summative assessment; (3) integration of instruction and assessment; and (4) integration of the portfolio as an innovative educational concept for formative and summative assessment purposes.

In Table 2, an example is given of the summative use of PPS. In this table, an example is presented of summative assessment scores that Jane Bond, a fictitious learner at the police academy, gained on six conventional assessment tasks in two task classes respectively with regard to the competence 'Supervise a public domain'. This learner had already practised the supervision of a public domain in several learning tasks with support before the assessment tasks. Three main requirements that need to be taken into account when developing PPS are illustrated by means

Table 2. Overview of fictitious assessment scores in PPS.

Student: Jane Bond														
Competence: supervise a public domain														
				Vertical standards (for eight criteria)								Horizontal standards (over scored criteria)		
Task class	No.	Assessment task[b]		4.7[a]	3.5	2.5	3.5	3.5	3.5	4.0	4.5			
				Score per criterion (maximum score = 6 for each criterion)								Horizontal standard[c]	Mean score	Decision[d]
1	1.5	SJT		3		3	3	4	2		3			
			Mean score	3.0		3.0	3.0	4.0	2.0		3.0	3.70	3.0	–
			Decision	–		+	–	+	–		–			
1	1.6	WST		5	2	1	4	3	4	2				
			Mean score	4.0	2.0	2.0	3.5	3.5	3.0	2.0	3.0	3.60	2.9	–
			Decision	–	–	–	+	+	–	–	–			
1	1.7	POJ		6	5	5	5	5	6	6	5			
			Mean score	4.7	3.5	3.0	4.0	4.0	4.0	4.0	4.0	3.71	3.9	+
			Decision	+	+	+	+	+	+	+	–			
2	2.6	SJT		3	2	4		3		3	3			
			Mean score	3.0	2.0	4.0		3.0		3.0	3.0	3.78	3.0	–
			Decision	–	–	+		–		–	–			
2	2.7	WST		5	4	3	2	4			5			
			Mean score	4.0	3.0	3.5	2.0	3.5		3.0	4.0	3.70	3.3	–
			Decision	–	–	+	–	+		–	–			
2	2.8	POJ		6	6	5	5	5	4	5	6			
			Mean score	4.7	4.0	4.0	3.5	4.0	4.0	4.0	4.7	3.71	4.1	+
			Decision	+	+	+	+	+	+	+	+			
3	3.5	WST	Etc.											

Notes: [a]This is the minimum score on the first criterion. [b]SJT = a Situational Judgement Test; WST = a Work Sample Test; POJ = a Performance On the Job assessment. [c]Mean of the measured vertical standards. [d]– = the learner fails on the competence level of this complexity level, more assessment tasks are required, + = the learner passes on the competence level of this complexity level, learner proceeds to the next complexity level.

of this example. These requirements are: (1) a mix of assessment tasks; (2) a standard set of assessment criteria; and (3) horizontal and vertical evaluation.

PPS requirement 1: mix of assessment tasks to assure quality

A first requirement of PPS is that a range of assessment tasks is necessary to gather reliable and valid information about a learner's competence. The standard approaches to reliability and validity are derived from a psychometric approach (Johnston 2004), while in competence-oriented learning contexts a shift is observed from psychometric to edumetric criteria for the quality of assessment scores (Dierick, van de Watering, and Mujtjens 2002). There is more attention to criteria such as accuracy of the scores, the cognitive complexity, the authenticity and transparency of the assessments, and the fairness in assessment (Baartman, Bastiaens, and Kirschner 2006). Three important criteria to improve the quality of the assessments are specifically addressed in PPS: accuracy, generalisability, and extrapolation.

An assessment score 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 (De Groot 1969). Scores are never fully accurate. Each assessment score is the product of the true performance and an error. These errors can be caused by factors related to the person (internal errors, e.g. motivation, physical aspects) or the environment (external errors, e.g. the tasks, the conditions of the environment, the assessors, the procedure). In particular the inter-rater reliability between assessors in assessment is rather low and needs improvement (see for example Straetmans 1998; Lunz, Wright, and Linacre 1990). Arriving at agreement over outcomes of complex assessment tasks included in portfolios is still one of the main concerns (Johnston 2004). Working with one assessor is undesirable, but working with more assessors is often impracticable. A solution for this problem is the training of assessors, where support is given in the construction of detailed assessment protocols in which the assessment criteria are defined specifically and unambiguously (De Graaff 1993; Straetmans 1998). Several studies report positive effects of discussions between assessors before grading (Nystrand, Cohen, and Dowling 1993; Heller, Sheingol, and Myford 1998; Pitts et al. 2001). The drawback of these detailed protocols, however, may be that the assessment is too analytical. Nystrand, Cohen, and Dowling (1993) and Pitts et al. (2001) investigated whether it is preferable to have a holistic approach in authentic assessment. When competences are assessed through a task that requires the learners to integrate them, 'holistic' or 'integrated' assessment is required. This form of assessment requires observation of performance in which a number of skills are interrelated and observed as a whole. This definition is in line with the assumptions of the 4C/ID model.

The assessment of competences also implies more than one observed performance. The learner has to perform a similar type of task in a variety of situations under the same conditions. Studies generally conclude that the *generalisability* to performances in similar tasks is limited (Linn, Baker, and Dunbar 1991). The main reason for this finding is that the assessment tasks in current curricula 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 to choose a variety of assessment tasks that represent a certain level of authenticity (Gulikers, Bastiaens, and Kirschner 2004). Following the assumptions of the 4C/ID model, costs and time can be reduced by reorganising the learning and assessment tasks according to the whole-task approach.

Extrapolation implies that the attained score reflects the performance level that the learner would achieve in a real working situation. Sometimes this is not a problem, because the assessment task does not deviate from the task in the real situation. But often it is. For example when

the performance task is too expensive (launch of a Patriot missile), too dangerous (defuse a bomb), or when the situation is unlikely to occur in real life (the arrest of an armed criminal in a shopping centre). The level of authenticity is thus defined by its degree of resemblance to the criterion situation (Gulikers, Bastiaens, and Kirschner 2004). 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 put heavy demands on the organisation of authentic assessments. Each type of assessment task has a weak link in the quality chain that links the performance of the learner and the conclusion regarding the competence in a particular context (Crooks, Kane, and Cohen 1996). The quality criteria are also problematic in the sense that optimising one criterion leads to an impairment of another criterion (Kane 1992). Therefore, it is important to choose a *mix of assessment methods* (Straetmans and Sanders 2001). In Table 2, some examples of assessment methods are applied to gather pieces of evidence for the mastery of the competence (see the column 'Assessment Task').

The Situational Judgment Test (SJT) is a hands-off assessment (Straetmans 1998). The learner is confronted with a realistic description of a situation he or she may encounter in professional life. The learner has to choose from a number of possible courses of action. Both the description and the courses of action can be presented in an authentic way, for example with a simulation. In spite of this, only a few people would be willing to take a summative decision based on an SJT performance only. This test particularly measures professional knowledge, i.e. knowing what to do in a so-called critical situation. In combination with other methods, however, the SJT is a valuable method in gathering evidence for competence development. In a relatively short period of time, a large number of critical situations can be presented to the learner. The learner has to demonstrate that she/he is proficient in taking decisions in a number of different situations (transfer).

In a Work Sample Test (WST) the learner is asked to perform a task under simulated conditions. The task is carried out in an authentic environment with all the sources that a professional has at his or her disposal. The main difference with 'real life' is that specific behaviours are provoked by giving certain stimuli. An immediate – and possibly negative – consequence is that the learner is more alert than in a less provoked situation. With regard to the competence 'Supervise a public domain', a work sample test could be the following:

> A trainee police officer was told that the railway station master at the local station asked for more supervision in the area near the station, because there are robberies on a regular basis. The trainee police officer is ordered to supervise the area near the station in the twilight. When he arrives at the station he spots a drifter looking for a place to sleep. The drifter is an actor, who is told to act aggressively. He is also the ward of a group of youngsters (also actors) who use the square in front of the station as a skating area.

A characteristic of a competent person is that she/he autonomously recognises the signals of a problem and subsequently acts towards an adequate solution. The predictive value of the work sample test regarding professional practice is dependent on the provocation of the desired behaviour. The predictive value of the work sample test decreases when the desired behaviour is explicitly provoked.

In a Performance On the Job assessment (POJ) the learner demonstrates his/her competence in an authentic setting. The only difference is that the learner is observed and assessed. Not surprisingly, this method is the best predictor for future practice of the learner. A restriction of the POJ is that it is not always feasible to organise a setting that represents the complexity level of the task class.

PPS requirement 2: standard set of assessment criteria

During an educational programme, a learner gathers evidence for specific competences. These pieces of evidence are used to take a decision about the learner's competence level. Evidence is gathered through a number of assessment tasks (see first PPS requirement), carried out in several contexts at several times, and assessed by multiple assessors. But when do you conclude that a learner has mastered a competence at a specific complexity level? Educational programmes designed according to the 4C/ID model always specify the standards for all criteria of performance in a certain task class. This satisfies the second requirement of PPS, namely that one prespecified set of criteria for acceptable performance is available to assess a learner on each learning task (see the top row in Table 2, where performance is judged on eight criteria). The standards are derived from the performance criteria, and specify minimum requirements on aspects of the competence. Pursuing our example of the police officer, examples of criteria of the constituent skill 'manage conflicts between two or more persons' are: (1) linguistic usage is attuned to the receiver; (2) objections are adequately tackled; (3) the non-verbal communication is effective and correct; and (4) the way of acting leads to the desired behaviour. Usually, content experts determine the minimal score (i.e. the standard) for all relevant performance criteria that the learner has to reach on each criterion to achieve an acceptable competence level. This minimal score is called the 'vertical standard' (see Table 2; the minimal score for criterion three for example is 2.5).

Which criteria are scored is dependent on the assessment task. One of the guidelines of PPS is that an assessor carefully observes the learner during the assessment task. Based on the observation the assessor(s) can decide which criteria are scored. In the fictitious portfolio of Jane Bond (see Table 2) it appears that the first piece of evidence in the first task class (SJT) is scored on six criteria, the second (WST) on seven criteria, and that in the performance on the job assessment (POJ) all criteria could be observed and scored.

PPS requirement 3: horizontal and vertical evaluation

Table 2 indicates that at each point in time, when the assessment results of a new assessment task are added to the scoring system, decisions can be based on a 'horizontal evaluation' and a 'vertical evaluation'. The horizontal evaluation indicates to what degree the standards are met for a learner's overall performance; it reflects the learner's level of mastery of the whole competence in a specific task class. The vertical evaluation indicates to what degree the standards are met for one particular criterion of a learner's performance, assessed by different assessment tasks; it reflects the learner's level of mastery of only one criterion of the competence. We illustrate this with the scores presented in Table 2. Based on the performance of the learner on the first assessment task in the first task class (1.5: a SJT), a mean horizontal score of 3.0 is computed. This is below the horizontal standard of 3.70, which is based on the mean of all observed vertical standards. Furthermore, the vertical evaluation results indicate that four criteria are below standard, and that in the next assessment task the criteria two and seven should be emphasised because these criteria could not be scored in the first assessment task. Therefore, a second assessment task (1.6: a WST) will be given to provide additional evidence of competence mastery on the first complexity level. The scores yielded on this task lead to a mean horizontal score of 2.9, which is still below the standard. Therefore, a third assessment task will be given as the next assessment task (1.7: a POJ). Table 2 shows that in this assessment task all criteria are scored, and that the mean horizontal score (3.9) is now above standard. The vertical evaluation results show that only one criterion is still below standard. Based on these PPS results, it is decided here that the learner is competent on the complexity level of this task class and may progress to the next complexity level. The process described above repeats itself until the learner successfully

performs the final assessment task in the most difficult task class. This task represents the competence level of a starting professional.

Flexibility in learning: the surplus value of PPS

Contemporary educational programmes aim at increased flexibility by adapting the learning path to individual learner needs. For several years, it has been possible to observe a movement that involves a transition from supply-driven education where the supply is 'absorbed' by the learner, to demand-driven education in which learners are challenged to prove in their own way that they are competent (Kirschner and Valcke 1994). Demand-driven education highlights the learner's role and needs while the teacher becomes a manager, mediator and motivator of learning (Kirschner, Vacke, and Sluijsmans 1999). It enables task-driven learning in which learners consult instructional materials in order to acquire relevant competences. Some learners have competences acquired elsewhere that should be taken into account, and some learners are better able to acquire new competences and therefore need less practice and guidance than other learners. This flexible learning allows learners to follow a learning path at an optimal pace and integrate the learning process into their personal development schedule.

Whereas 'traditional' assessments are more or less fixed in curricula and seen as control processes at the end of a study programme, PPS provides opportunities to integrate assessment in each learning task which enables adaptation of the personal learning path to the needs of the learner. In such a flexible curriculum, not all learners receive the same sequence of learning tasks (i.e. one educational programme for all), but each learner receives his or her own sequence of learning tasks that is adapted to individual needs, progress and preferences. Assessment is critical to the selection of the next suitable learning task. PPS supports reliable and valid assessments of performances whereby the output of PPS is used to interpret learning outcomes. At one extreme, it can be the system that assesses a learner's progress and selects the next learning task for the learner to work on. At the other extreme, it is the self-regulated learner who continuously self-assesses his or her progress and selects the next learning task from all available tasks. In between, the assessment itself and the interpretation of assessment results will be a shared responsibility of the system or teacher and the learner. The responsibility of the learners may increase as they further develop the self-regulation skills that are necessary to select suitable learning tasks, including not only self-assessment skills but also orienting skills (what could I learn from this task?), planning skills (how much time and effort would I need to invest in this task?), monitoring skills (did I learn enough to stop working on this task?), and so on. In this process, negotiation between teachers and learners is important, since teachers and learners may interpret relevant aspects of task performance and learning tasks differently (see Bjork 1999), which can lead to undesirable differences in the learning effectiveness.

With regard to flexible learning, the responsibility for assessment is gradually handed over to the learner (Sluijsmans 2002), who takes responsibility for arranging a personal learning path. The development of self-regulation skills, which is increasingly seen as an important goal of education, includes the individual ability to select learning tasks that best help to reach educational and personal goals. Integrating forms of self- and peer assessment – whereby learners assess their own competences or those of peers – proved to be valuable learning tools for learners in terms of improved task performance and self-regulatory skills (Sluijsmans, Dochy, and Moerkerke 1999).

Discussion

In this article, we have described how authentic assessment can be integrated in CBL in vocational education with a specific focus on a new approach for the storage of assessment results:

Protocol Portfolio Scoring. Characteristic of this approach is the design of integrative learning and assessment tasks. According to the 4C/ID model, these tasks are organised from a task class with relatively simple tasks to the task class with tasks that reflect the desired performance of a starting professional. Within each task class learners carry out tasks that are equally difficult, but differ on dimensions on which tasks in real life also differ. Moreover, learners receive much support during their first tasks in a task class. The final tasks in a task class are performed without any support: the learner has to perform these autonomously. These tasks are suitable for summative assessment.

The 4C/ID approach contrasts sharply with other approaches in vocational education. First, the model departs from 'whole' integrative tasks. These tasks set up the backbone of the curriculum and are essential for the intended integration of knowledge, skills and attitudes. This integration is necessary to enable a natural transfer from the educational area to the professional area. Implementation of the 4C/ID approach will have major consequences for a curriculum that is built up from separate content domains, because these domains need to be integrated and rearranged in such a way that they optimally support learners' work on integrative learning tasks.

A second distinction is the continuous variation between theory and practice. The old adage 'theory first, then practice' is relevant here. The information that supports learners during working on the learning tasks is presented or discussed with the learners before a task class and/or consulted by the learners during the task class. The new, additional information that is needed to perform the tasks in a task class with a higher level of complexity is presented in the next task class, so that theory and practice are intertwined. The just-in-time information that is related to recurrent aspects of the competence is preferably presented during work on the task (e.g. a coach, a 'job aid' or an Electronic Performance Support System). A third distinction concerns the meaning of study progress and study success. The belief that study progress can be described as an accumulation of credit points is no longer a fact. Study progress implies the process of demonstrating competence over and over again, in contexts in which the complexity increases. Only this guarantees gradual competence development towards the level of the starting professional. Training based on the 4C/ID model, in which learners work on authentic task, is also more challenging and motivating.

The proposed method for making and assessing a portfolio based on results of assessments tasks, Protocol Portfolio Scoring, constitutes a new vision for learners' study progress. PPS acknowledges the problems that are inherent to assessment. The issue of validity is addressed by choosing a variety of whole-task assessment methods (the 'method mix'), whereby the quality criteria for assessment are warranted. Each new piece of evidence provides information about the competence, translated in terms of scores on a set of assessment criteria. Decisions regarding the level of competence are based on vertical and horizontal evaluation of the scores gained. PPS intends to build up and assess a portfolio in a systematic and solid way, in order to reach accurate decisions about the competence level of learners. Furthermore, PPS can be of additional value in the context of a demand-driven paradigm in which tasks and assessments are adapted to the needs and requirements of learners. The increase in learner control implies more self-regulated learning and involvement of learners in the selection of assessment tasks and the interpretation of assessment results. Although active, self-regulated learning has proved to lead to higher learning outcomes (Boekaerts 1997; Zimmerman 2002), research also shows that in general learners have difficulty self-regulating their learning (Hofer, Yu, and Pintrich 1998; Peverly et al. 2003). It cannot be assumed beforehand that learners are capable of reliable self-assessment, which is underpinned by research on involvement of learners in assessment (Sluijsmans et al. 2004). The main question is how teachers can select learning tasks for learners who are not yet able to do this, how they can help learners to take more and more responsibility for selecting their own learning tasks, and how they can give proper advice and guidance.

In their search for flexible forms of learning in which 'demand-driven education' is becoming increasingly important, many institutes in vocational and higher education are currently moving to a 'supermarket model', in which learners have total freedom to select any learning task or course they like, at any point in time. Models as described in this article may help educational institutions to increase this flexibility of their educational programmes while at the same time maintaining high-quality assessment. The 4C/ID model and PPS both offer valuable guidelines, and the practical implications are straightforward and ambitious. However, the educational value of both approaches is still unsatisfactorily investigated. We also realise the immense effort that is needed for a successful PPS system. To determine whether the approach is suitable and feasible on a large scale and which elements from the 4C/ID model and the PPS are functioning adequately, further research is necessary.

With the implementation of PPS, where feedback and reflection by the students themselves could become central processes, it can be expected that students will experience the assessment situations as more supportive for their development than in teacher-controlled assessment situations. Moreover, it can be expected that students perceive the assessment setting as being in control over their assessment. Finally, by discussing assessment criteria and standards, it is expected that PPS will convey competence to students. In sum, it is expected that in school settings where PPS is implemented, students' perception of control and competence will increase intrinsic motivation and in turn the quality of their learning.

References

Arter, J. 1996. Using assessment as a tool for learning. In *Student performance assessment in an era of restructuring,* ed. R. Blum and J. Arter, 1–6. Alexandria, VA: Association for Supervision and Curriculum Development.

Alderson, J.C., and D. Wall. 1993. Does washback exist? *Applied Linguistics* 14: 115–129.

Baartman, L.K.J., T.J. Bastiaens, and P.A. Kirschner. 2006. The wheel of competency assessment: Presenting quality criteria for Competency Assessment Programmes. *Studies in Educational Evaluation* 32: 153–177.

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 *Optimising new modes of assessment: In search of qualities and standards,* ed. M. Segers, F. Dochy, and E. Cascallar, 13–36. Dordrecht: Kluwer.

Bjork, R.A. 1999. Assessing our own competence: Heuristics and illusions. In *Attention and Performance XVII: Cognitive regulation of performance: Interaction of theory and application,* ed. D. Gopher and A. Koriat, 435–459. Cambridge, MA: MIT Press.

Black, P., and D. Wiliam. 1998. Assessment and classroom learning. *Assessment in Education: Principles, Policy and Practice* 5, no. 1: 7–74.

Boekaerts, M. 1997. Self-regulated learning: A new concept embraced by researchers, policy makers, educators, teachers, and students. *Learning and Instruction* 7: 161–186.

Borko, H., P. Michalec, M. Timmons, and J. Siddle. 1997. Student teaching portfolios: A tool for promoting reflective practice. *Journal of Teacher Education* 48: 345–357.

Crooks, T.J., M.T. Kane, and A.S. Cohen. 1996. Threats to the valid use of assessments. *Assessment in Education* 3, no. 3: 265–285.

De Graaff, M. J.C. 1993. Managers as assessors. In *Assessment centers: An open book,* ed. P.G.W. Jansen and F. de Jong, 62–73. Utrecht: Spectrum.

De Groot, A.D. 1969. *Methodology. Foundations of inference and research in the behavioral sciences.* The Hague/Paris: Mounton & Co.

Dierick, S., G. van de Watering, and A. Muijtjens. 2002. De actuele kwaliteit van assessment: ontwikkelingen in de edumetrie [Current quality in assessment: Developments in edumetrics]. In *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 competence-based education], ed. F. Dochy, L. Heylen, and H. van de Mosselaer, 91–122. Utrecht: Lemma.

Dochy, F.J.R.C., and L. McDowell. 1997. Assessment as a tool for learning. *Studies in Educational Evaluation* 23: 279–298.

Frederiksen, N. 1984. The real test bias, influences of testing on teaching and learning. *American Psychologist* 39: 193–202.

Heller, J.I., K. Sheingol, and C.M. Myford. 1998. Reasoning about evidence in portfolios: Cognitive foundations for valid and reliable assessment. *Educational Assessment* 5: 5–40.

Hofer, B.K., S.L. Yu, and P.R. Pintrich. 1998. Teaching college students to be self–regulated learners. In *Self-regulated learning: From teaching to self-reflective practice,* ed. D.H. Schunk and B.J. Zimmerman, 57–85. New York: Guilford Press.

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

Gulikers, J., Th. Bastiaens, and P. Kirschner. 2004. A five-dimensional framework for authentic assessment. *Educational Technology Research and Development* 52: 67–85.

Gulikers, J., Th. Bastiaens, and P. Kirschner. 2006. Authentic assessment, student and teacher perceptions: The practical value of the five dimensional framework. *Journal of Vocational Education and Training* 58, no. 3: 337–357.

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

Kane, M.T. 1992. The assessment of professional competence. *Evaluation and the Health Professions* 15: 163–182.

Kirschner, P., and M. 1994 Valcke.. From supply-driven to demand-driven education: new conceptions and the role of ICT therein. *Computer in Human Services* 10: 31–53.

Kirschner, P., M. Valcke, and D. Sluijsmans. 1999. Design and development of third generation distance learning materials: From an industrial second generation approach towards realizing third generation distance education. In *Design approaches and tools in education and training,* ed. J. van den Akker, R.M. Branch, K. Gustavson, N. Nieveen, and T.J. Plomp, 81–93. ICO: Kluwer Academic Publishers.

Le Mahieu, P., D. Gitomer, and J. Eresh. 1995. Portfolios in large-scale assessment: difficult but not impossible. *Educational Measurement: Issues and Practice* 14: 11–10.

Linn, R.L., E. L. Baker, and S.B. Dunbar. 1991. Complex, performance-based assessment: Expectations and validation criteria. *Educational Researcher* 20, no. 8: 15–21.

Lunz, M.E., B. Wright, and M. Linacre. 1990. Measuring the impact of judge severity on examination scores. *Applied Measurement in Education* 3, no. 4: 331–350.

Meyer, C. 1992. What's the difference between authentic and performance assessment? *Educational Leadership* 49: 39–40.

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

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

Peverly, S.T., K.E. Brobst, M. Graham, and R. Shaw. 2003. College adults are not good at self-regulation: A study on the relationship of self-regulation, note taking, and test taking. *Journal of Educational Psychology* 95: 335–346.

Pitts, J., C. Coles, and P. Thomas. 2001. Enhancing reliability in portfolio assessment: 'Shaping' the portfolio. *Medical Teacher* 23: 351–355.

Reckase, M.D. 1995. Portfolio assessment: A theoretical estimate of scoring reliability. *Educational measurement: Issues and Practice* 14: 12–14, 31.

Sluijsmans, D.M.A. 2002. Student involvement in assessment: The training of peer assessment skills. Unpublished doctoral diss., Open University of the Netherlands, Heerlen.

Sluijsmans, D.M.A., S. Brand-Gruwel, J. van Merriënboer, and R. Martens. 2004. Redesigning education for training peer assessment skills in teacher education. *Innovations in Education and Training International* 41, no. 1: 59–78.

Sluijsmans, D.M.A., F. Dochy, and G. Moerkerke. 1999. Creating a learning environment by using self- peer- and co-assessment. *Learning Environments Research* 1: 293–319.

Smith, K., and H. Tillema. 1998. Evaluating portfolio use as a learning tool for professionals. *Scandinavian Journal of Educational Research* 41: 193–205.

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

Straetmans, G.J.J.M. 1998. Toetsing van competenties [Competence assessment]. In *Handboek Effectief Opleiden [Handbook for Effective Education] (9.1–3.01–3.36),* ed. P.W.J. Schramade. 's-Gravenhage: Delwel Uitgeverij B.V.

Straetmans, G.J.J.M., and P.F. Sanders. 2001. *Beoordelen van competenties van docenten* [Assessing teachers' competences]. Den Haag: Programmamanagement EPS/HBO-raad.

Straetmans, G., D. Sluijsmans, B. Bolhuis, and J. van Merriënboer. 2003. Integratie van instructie en assessment in competentiegericht onderwijs [Integration of instruction and assessment in competence-based education]. *Tijdschrift voor Hoger Onderwijs* 3: 171–197.

Sweller, J., J.J.G. van Merriënboer, and F.G.W.C. Paas. 1998. Cognitive architecture and instructional design. *Educational Psychology Review* 10: 251–296.

Tillema, H.H., J.W.M. Kessels, and F. Meijers. 2001. 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 Merriënboer, J.J.G. 1997. *Training complex cognitive skills: A four-component instructional design model for technical training.* Englewood Cliffs, NJ: Educational Technology Publications.

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

Wade, R.C., and D.B. Yarbrough. 1996. Portfolios: A tool for reflective thinking in teacher education? *Teaching and Teacher Education* 12: 63–79.

Wiggins, G. 1989. Teaching to the (authentic) test. *Educational Leadership* 46, no. 7: 41–47.

Zimmerman, B.J. 2002. Achieving academic excellence: A self-regulatory perspective. In *Pursuit of excellence through education,* ed. M. Ferrari, 85–110. Mahwah, NJ: Lawrence Erlbaum.