Design and evaluation of video portfolios : reliability, generalizability, and validity of an authentic performance assessment for teachers

Bakker, M.E.J.

Chapter 2

Video portfolios: The development and practical utility of an authentic teacher assessment procedure[1]

Abstract

This chapter reports on the design and practical utility of an authentic assessment procedure that can be used for assessing teachers' coaching competence in the context of senior secondary vocational education. The aim was to determine to what extent assessors are able to cope with the assessment procedure designed, and to explore how assessors can be supported in making valid interpretations and judgments. Video recordings of teachers' coaching performance in the classroom are the main elements of the assessment procedure constructed. Additional data sources were included that provide information on the context of the videotaped coaching situations. This combination of video recordings and context information is called a 'video portfolio'. Six trained assessors scored three video portfolios. The scores they assigned were collected and the interrater agreement was determined. After the video portfolios had been scored, the assessors were interviewed about their experiences of scoring and judging them. The overall conclusion is that assessors seem to be reasonably capable of using the scoring procedure, and that it yields relatively comparable judgments. The assessors indicated that it is necessary to be trained in using the assessment procedure, and that following this procedure takes a lot of energy. Particularly mastering the scoring method takes much time.

2.1 Introduction

The last two decades, new forms of teacher assessment have been developed and used. These new forms of assessment, often referred to as 'performance assessment' and 'authentic assessments', reflect a shift in assessment purposes and conceptions of teaching (Darling-Hammond & Snyder, 2000; Haertel, 1991; Gipps, 1994). New views on teacher assessment place more emphasis on the formative function of assessment,

[1] This chapter has been submitted in adapted form as:
Bakker, M., Roelofs, E., Beijaard, D., Sanders, P., Tigelaar, D., & Verloop, N. Video portfolios: The development and practical utility of an authentic teacher assessment procedure.

in which assessment results are used for teachers' further professional development. To ensure that assessment tasks make up a meaningful learning experience, it is argued that they should be authentic and realistic to teachers who are being assessed (Uhlenbeck, Verloop, & Beijaard, 2002). In the new conception of teaching, teaching is recognized as a complex activity that is highly contextual and personal (Darling-Hammond & Snyder, 2000; Dwyer, 1994). To ensure that assessments reflect these conceptions, assessments should be authentic and emphasize the assessment of actual teaching performance in complex everyday conditions. Based on the changing conceptions of teaching, new methodologies have emerged for authentic assessment, like teacher work samples (Girod, 2002; Salzman, Denner, Bangert, & Harris, 2001; Schalock, Schalock, & Girod, 1997), structured types of teacher portfolios (Barton & Collins, 1993, 1997; Seldin, 1991; Wade & Yarbrough, 1996), and, more specific, video portfolios (Frederiksen, Sipusic, Sherin, & Wolfe, 1998). In these methods the focus is on collecting and judging evidence using a deliberately chosen sample of instructional activities based on a curricular unit. The various forms of evidence all refer to the same set of instructional situations, which are deliberately set out to attain learning objectives (Girod, 2002).

These new purposes, conceptions, and methods of teaching and teacher assessment have consequences for the design of authentic teacher assessments. A common knowledge base is gradually emerging of what constitutes valid, reliable, and educative authentic teacher assessments. Frequently cited design principles and underlying notions for authentic teacher performance assessments are listed in Table 2.1.

Table 2.1 Design principles for authentic performance assessments

Design principles
1. The scoring method used by assessors should be systematic and transparent In general, the scoring methods in authentic assessments are rather complex. To reduce the risk of invalid and unreliable judgments, it is common to use a systematic and transparent scoring method. In addition, assessors are usually trained to use the scoring method consistently (Gipps, 1994; Linn, Baker, & Dunbar, 1991).

Table 2.1 Design principles for authentic performance assessments (Continued)

Design principles
2. Criteria and performance levels should include theoretical perspectives on competent teaching as well as practice-based perspectives In order to obtain representative criteria and performance levels for judging competent teaching, theoretical as well as practice-based perspectives should be included in the criteria and standards (Uhlenbeck, 2002). 3. Criteria and performance levels should describe essential aspects of professional performance in terms of what professional teachers should know and be able to do A major issue in formulating criteria and performance levels is choosing the appropriate level of specificity of these key aspects of professional performance. If the criteria or performance levels are formulated too broadly/generally, then it is difficult for assessors to apply these criteria and performance levels consistently. If the criteria and performance levels are formulated too narrowly/specifically, then there is the risk of getting lost in specifics and the essence of teaching is missed (Dwyer, 1993; Kagan, 1990). 4. Criteria and performance levels should not favour any style of teaching Criteria and performance levels used in teacher assessment should specify what aspects of teaching will be assessed and not how teachers should carry them out (Darling-Hammond & Snyder, 2000; Dwyer, 1998; Gipps, 1994; Haertel, 1991; Uhlenbeck, 2002). 5. Multiple methods should be used to cover different aspects of teaching Not only relevant performance (acting) should be assessed, but also relevant knowledge and decisions that underlie performance (Beijaard & Verloop, 1996; Dwyer, 1998; Uhlenbeck, 2002). 6. The assessment should take place in a context that closely resembles the actual teaching context It is recognized that all teaching and learning is shaped by the context in which it occurs. Factors like grade level, subject, students' ability, and school policy largely determine what approaches to teaching will be effective, and it is, therefore, important to include the context in assessment tasks (Darling-Hammond & Snyder, 2000; Dwyer, 1998; Gipps, 1994; Haertel, 1991; Uhlenbeck, 2002). 7. Assessment tasks should reflect the complexity of teaching Teaching involves immediate and adequate decision-making and acting in a specific situation, in which a teacher has to take many variables into account. Assessment tasks should include this immediate decision-making and acting in a specific context (Darling-Hammond & Snyder, 2000; Dwyer, 1998; Gipps, 1994; Haertel, 1991; Uhlenbeck, 2002).

In this chapter we present the design of an authentic assessment procedure in which the principles listed in Table 2.1 were taken as starting point. This assessment procedure was aimed at assessing teachers' coaching competence in the context of senior secondary vocational education. As a result of the implementation of competence-based teaching in the Netherlands, coaching has become an important domain of teacher competence. It is expected that teachers who take on a coaching role will contribute to self-regulated and independent learning of the learners, one of the central aims of competence-based learning in vocational education (Moerkamp, De Bruijn, Van der Kuip, Onstenk, & Voncken, 2000; Onstenk, 2000). In this relatively new learning environment, teachers are supposed to coach students who work collaboratively in small groups on complex, job-related tasks.

Different kinds of evidence for assessing teachers' coaching competence were gathered. Inspired by the work of Frederiksen, Sipusic, Sherin, and Wolfe (1998), a procedure for video portfolio assessment was set up. The main evidence collected consisted of video recordings of teachers' coaching performance in the classroom. The videos were collected in a systematic way, and additional data sources were added that outlined the context in which the coaching took place. This collection of documented video registrations and context information is referred to as a video portfolio. Video portfolios are supposed to provide assessors with a structured and well-documented collection of evidence with regard to the coaching performance. In this study, researchers constructed the video portfolios.

An authentic assessment procedure based on video portfolios consists of rich, qualitative information that requires interpretations and judgments of assessors to determine what it means. The validity of their interpretations and judgments largely determines the quality of the assessment. However, it is not easy to interpret and judge qualitative data in a consistent, objective, and comparable way (Gipps, 1994; Moss, 1994). The aim of this study was to explore to what extent assessors were able to apply the designed procedure for assessing video portfolios, and to explore which aspects of the procedure supported or hindered the assessors in making valid interpretations and judgments.
The specific research questions were the following:

- To what extent do assessors arrive at corresponding scores for video portfolios when judging them using the designed scoring procedure?

- Which aspects of the video portfolio assessment procedure stimulate or hinder the assessors in making valid interpretations and judgments?

The development of the assessment procedure is first described. Attention is given to the domain of competence that was assessed; the criteria and performance levels designed for the assessment; the kinds of evidence that were gathered, structured, and documented for the assessment; how the evidence was to be scored and judged by the assessors; and in what way the assessors were trained in applying the scoring method.

2.2 Development of the assessment procedure

In designing the assessment procedure, we started by defining the domain of competence to be assessed. The video portfolios were to be used to assess teachers' coaching competence in senior secondary vocational education. In the context of this innovation, a specific project was started called the 'MTS+ project'. In the MTS+ project, a specific learning environment was developed to foster self-regulated learning in the context of technical studies. Teachers' task in this context was to coach students who work collaboratively on complex tasks. Relevant domain specific knowledge and (to some extent) skills related to building and construction techniques had to be applied while students worked on these tasks. The students were asked to carry out authentic tasks such as designing holiday homes and building a dam. The tasks were relatively large projects; students worked on a single task for approximately four weeks. In order to accomplish a task, students were expected to carry out various learning activities. In this learning context, teachers were expected, first, to coach students in performing learning activities that they could not (yet) carry out on their own, and, second, to coach students in developing realistic perceptions of professional thinking and acting in practice.

The coaching of students in the new learning environment was elaborated into an interpretive framework, which was used for scoring and judging the video portfolios. This interpretive framework reflected all relevant aspects of coaching competence, performance criteria, and scoring instructions which would enable assessors to judge teachers' performance using the video portfolios.

The main purpose of the interpretive framework was to prevent assessors from scoring and judging video portfolios according to their own criteria as much as possible. It is known from the literature that assessors, while assessing, use schemata in understanding and predicting respondents' behavior (DeNisi, Cafferty, & Meglino, 1984; Feldman, 1981; Landy & Farr, 1980). The schemata are comparable to personal constructs (Kelly, 1955), and are used to organize and simplify information. The schemata work like filters, causing assessors to look selectively at information and interpret it according to their own constructs (Van der Schaaf, Stokking, & Verloop, 2005). In providing an interpretive framework, the assessors were urged to score and judge the video portfolios according to the constructs in the framework.

2.2.1 Defining teaching interventions that can be marked as coaching interventions

The first step in designing criteria and performance levels for assessing teachers' coaching competence was to formulate interventions that could be marked as coaching interventions. This part of the interpretive framework was meant to support assessors in identifying and judging coaching interventions out of all teaching interventions taking place during the performance. The design principles 1 and 2 as presented in Table 2.1 were the starting point in defining coaching activities. In accordance with principle 1, coaching interventions were defined based on the findings of a literature study; these coaching interventions were adjusted and refined to suit the specific context of MTS+, so that theoretical as well as practice-based perspectives were represented in the framework. In order to conform with design principle 2, the goal was to capture only essential coaching interventions in the framework using a literature study and observations in practice aimed at extracting interventions commonly used in coaching situations.

Theoretical perspective on coaching interventions

From a theoretical point of view, coaching can be described as stimulating and supporting self-regulated learning (Boekaerts, 1999; Boekaerts, & Simons, 1995; Bolhuis, 2000; Butler & Winne, 1995). Typical coaching interventions that can be used to stimulate and support such learning are asking questions and providing feedback on learning activities employed by students. By asking questions and providing feedback, the teacher makes students aware of their learning activities and provides them with information about the adequacy, efficiency, and effectiveness of (performed) learning activities (Boekaerts & Simons, 1995; Butler & Winne, 1995). Students can use this

information to direct and regulate new learning activities. Providing clues, hints, advice, and examples also constitutes relevant coaching interventions (Boekaerts & Simons, 1995; Butler & Winne, 1995; Winne & Hadwin, 1998). Such feedback can be effective, for example, when students do not know how to continue their tasks or to find out where they made mistakes.

Coaching interventions are used to stimulate and support cognitive, meta-cognitive, affective learning activities, and learning activities concerning collaborative learning (Perry, 1998; Perry, Phillips, & Dowler, 2004; Shuell, 1993; Vermunt & Verloop, 1999). Cognitive learning activities concern processing activities that students use to process subject matter and that lead to learning outcomes in terms of changes in students' knowledge base and skills. Affective learning activities pertain to coping with emotions that arise during learning and that lead to a mood that fosters or impairs the progress of the learning process. Meta-cognitive regulation activities are thinking activities that students use to decide on learning contents, to exert control over their processing and affective activities, and to steer the course and outcomes of their learning (Vermunt & Verloop, 1999). Learning activities concerning collaborative learning concern communication, coordination, and realization of a positive group climate (Johnson & Johnson, 1994; Slavin, 1990).

Practice-based perspective on coaching interventions

In line with design principle 1, a practice-based perspective on coaching was also included in the interpretive framework. Based on observations and interviews with teachers, the literature-based framework was adjusted and refined to the specific context of MTS+. Five teachers participating in the MTS+ project were observed for two hours each during their coaching conferences with students, and interviewed afterwards. Three teachers were randomly selected; the other two teachers were pointed out as 'best coaches of the technical studies unit' by the principal.

From the observations and interviews, it was found that teachers use questions and give feedback as coaching interventions in the MTS+ context. The learning activities derived from the literature were recognized in practice and could be classified into more specific learning activities, which we labeled as 'aspects of learning activities'. Descriptions of the aspects of learning activities and related examples of coaching interventions are included in Appendix 1.

2.2.2 Defining criteria and performance levels for competent coaching

The second part of the interpretive framework specifies criteria and performance levels to be used by assessors to judge the quality of the individual coaching interventions and the entire coaching performance. Design principles 1, 2, and 3 were the starting point for defining criteria and standards for competent coaching. This implies that in this part of the framework theoretical as well as practice-based perspectives on competent coaching should be included. Furthermore, in formulating criteria and levels of performance, only aspects that distinguish competent from less competent coaching should be represented in this part of the interpretive framework. Observations in practice and a literature study were used to track down these aspects of competent coaching. In accordance with design principle 3, the criteria and levels of performance were formulated in terms of what a teacher should achieve.

In defining competent coaching, a general model for teachers' competence developed by Roelofs and Sanders (2007) was used as a starting point. According to this model, teachers' competence is defined as the extent to which a teacher, as a professional, takes deliberate and appropriate decisions (based on personal knowledge, skills, conceptions, etc.) within a specific and complex professional context (students, subject matter, etc.), resulting in actions which contribute to desirable outcomes, all according to accepted professional standards (Roelofs & Sanders, 2007). This definition shows the important relationship between teachers' actions and desirable consequences for students. It shows that competent performance is always directed towards positive consequences for students. Based on this notion, criteria for competent coaching were defined in terms of positive consequences for students in the context of MTS+.

Practice-based perspective on competent coaching

The criteria and performance levels were based mainly on the two learning goals that students are supposed to achieve in MTS+: (a) students should improve in employing learning activities while working on complex tasks, and (b) students should develop realistic perceptions of professional thinking and acting in practice. Competent coaching in this context can be defined as supporting students in achieving these learning goals.

Theoretical perspective on competent coaching

In line with design principle 1, a theoretical perspective on competent coaching was also included in the criteria and performance levels. The literature study was done in order to elaborate on how teachers can support students in achieving these learning goals. This part of the interpretive framework is based on theories concerning process-oriented instruction (Vermunt & Verloop, 1999; Vermunt & Verschaffel, 2000) and cognitive apprenticeship (Collins, Brown, & Newman, 1989).

Teachers' coaching of students in employing learning activities (learning goal a) can be defined as competent coaching when the teachers use coaching interventions that provide students with opportunities to improve their learning activities. Competent coaches provide just enough support in order to enable students to make the step to the next higher level in employing a learning activity, which they couldn't have made on their own (Vygotski, 1978). As the performance of a learning activity improves, the support of the teacher decreases until the student can perform the learning activity independently; this is referred to in the literature as 'fading' (Collins, Brown, & Newman, 1989). When the teacher is capable of providing just enough support to accomplish improvements in employment of a learning activity, coaching is considered 'constructive' (Vermunt & Verloop, 1999). When a teacher provides too much or too little support, improvement in conducting learning activities is expected not to take place. In that case, coaching is considered to be 'non-constructive' (Vermunt & Verloop, 1999). The performance levels of constructive coaching are presented in Table 2.2.

Table 2.2 Performance levels for constructive coaching

Level 4 **Rapid growth**	The teacher uses interventions that lead to many opportunities for students to improve in conducting learning activities. And/or He/she misses practically no opportunity to support/stimulate students in improving learning activities.
Level 3 **Growth**	The teacher uses interventions that lead to opportunities for students to improve in conducting learning activities. And/or He/she misses some opportunities to support/stimulate students in improving learning activities.
Level 2 **Faltering growth**	The teacher uses interventions that occasionally lead to opportunities for students to improve in conducting learning activities. And/or He/she misses many opportunities to support/stimulate students in improving learning activities.
Level 1 **No growth**	The teacher uses no interventions that lead to opportunities for students to improve in conducting learning activities. And/or He/she misses almost every opportunity to support/stimulate students in improving learning activities.

Teachers' coaching of students in developing realistic perceptions of practice (learning goal b) can be defined as competent coaching when teachers use coaching interventions that provide students with opportunities to improve their perceptions of practice through 'practice-oriented' coaching. A coach who plans to establish practice-oriented coaching should refer to rules, norms, procedures, methods, and typical situations that are used or occur in practice (Brown & Campione, 1994; Lave, 1991). When a teacher neglects to refer to professional practice during coaching, it is expected that students do not get a proper chance to construct representative views of professional thinking and acting in practice. The performance levels of practice-oriented coaching are presented in Table 2.3.

Table 2.3 Performance levels for practice-oriented coaching

Level 4 Full-grown perception of practice	The teacher uses interventions that lead to many opportunities for students to construct realistic perceptions of professional thinking and acting in practice. And/or He/she misses practically no opportunities to stimulate students in constructing realistic perceptions of professional thinking and acting in practice.
Level 3 Representative perception of practice	The teacher uses interventions that lead to opportunities for students to construct realistic perceptions of professional thinking and acting in practice. And/or He/she misses some opportunities to stimulate students in constructing realistic perceptions of professional thinking and acting in practice.
Level 2 Fragmented perception of practice	The teacher uses interventions that occasionally lead to opportunities for students to construct realistic perceptions of professional thinking and acting in practice. And/or He/she misses some opportunities to stimulate students in constructing realistic perceptions of professional thinking and acting in practice.
Level 1 No perception of practice	The teacher uses no interventions that lead to opportunities for students to construct realistic perceptions of professional thinking and acting in practice. And/or He/she misses almost every opportunity to stimulate students in constructing realistic perceptions of professional thinking and acting in practice.

2.2.3 Content of the video portfolio

In order to cover all aspects of coaching and to provide assessors with a complete picture of teachers' coaching competence, a mix of evidence is needed (design principle 4). The assumption was that assessors are better capable of understanding and interpreting the coaching performance shown in the video episodes when they know about the context in which the coaching performance took place. Research has shown that especially understanding the performance is a first and important step in making valid interpretations with regard to the performance (Heller, Sheingold, & Myford, 1998; Schutz & Moss, 2004).

The decision of what evidence to include in the video portfolio was based on the definition of teachers' competence mentioned in section 2.2.2. Primary and secondary sources of evidence for teacher competence can be distinguished. The primary sources

of evidence consisted of video episodes that represented coaching performance. For this, teachers were filmed on the job while they held coaching conferences with a group of students. The video recordings represented authentic performance in an authentic context (design principles 5 and 6). Other sources of evidence were added: interviews with the teachers about the decisions underlying their actions; interviews with students about the impact of teachers' actions on their work; information about students' background; information about the task students worked on during a video episode; information about students' progress in completing the task; and information about teachers' background. Assessors were expected to examine all these primary sources when assessing a video portfolio. The secondary sources of evidence consisted of educational materials students used during video episodes and students' products discussed during video episodes. Assessors could use the secondary sources of evidence in assessing a video portfolio if they felt the need for this extra information.

Recording professional performance

The researchers constructed four video portfolios of four teachers (one female coach and three male coaches). The participating teachers coached first-year students in MTS+ (of the technical vocational studies unit). All coaches had one to two years' experience in coaching students working on complex tasks. Each coach was filmed within a period of four weeks in which students completed one such task.

Before the actual recording, test recordings were carried out to get teachers and students used to the presence of video cameras. In addition, recording equipment and the positions of the different cameras were tested. Three cameras were used, placed around the students and the teacher. During coaching, teachers wore a wireless microphone, and two microphones were placed on the tables. The students and teachers were filmed frontally.

Documentation of video episodes

After four weeks of recording, 32 coaching sessions had been filmed. Coaching sessions varied from 20 to 60 minutes. The first step in documenting video episodes concerning relevant teacher performance was to synchronize and mix the three separate films to make one film, using professional edit software. Special guidelines were used for editing the film. For instance, in case of feedback to a specific student

or on a specific student product, a close-up was used and in the event of rapid interaction, the group-shot was used.

After the film was mixed, video episodes representing professional performance were selected from the recorded coaching conferences and were marked using time marks. In this process, the following guideline for selecting a video episode was used: the video episode had to be a situation in which students needed support in conducting a specific learning activity to complete the complex task they were working on. Such situations were expected to provide valuable evidence of teachers' coaching competence. In addition, for all video episodes, a short summary was written of what happened during the video episode, including what learning activity or activities the teacher supported. Information on the progress of the students in completing the task was also included in the summary.

Interviewing teachers and students and collecting context information

Immediately after the coaching session, two researchers made an initial selection of situations that occurred during the session, based on notes they took during the coaching conference. This selection of situations was used as input for the interview about the teachers' underlying decisions that resulted in performance. The teachers were interviewed directly after the recording of the coaching session. The specific interview questions used in the interview are included in Table 2.4. To retrieve information about the perceived effects of coaching, the students were also interviewed about the selected situations. Directly after the coaching conference, one or two students who received the most coaching were selected for the interview. The specific interview questions used in the interview are included in Table 2.4. In addition to the video episodes and the interviews, information about the students' and teachers' backgrounds, and copies of the instructional materials for teacher and students, were collected. Only information was gathered that was expected to support assessors' understanding of the performance shown in the video recordings. The specific information gathered is listed in Table 2.4.

Finalizing the video portfolio

The researchers selected marked video episodes for each teacher for inclusion in their video portfolios. It is known from the literature that assessors form a pattern of the data in a portfolio (Moss, Schutz, & Collins, 1998; Schutz, & Moss, 2004). A total of

ten episodes were selected for each teacher, because it was expected that ten video episodes and the corresponding context information would provide assessors with enough data to form a pattern with regard to teachers' coaching competence. Furthermore, it was expected that assessors would be capable of scoring ten video episodes within a reasonable length of time. Two further selection criteria were used. The selected set of video episodes should equally represent:

- four weeks of filming during which students were coached;
- different learning activities coached in MTS+.

All components of the video portfolio are summarized in Table 2.4. To arrange all the elements of a video portfolio in an orderly fashion, all evidence from the different sources was organized in a multimedia environment. An existing multimedia environment, MILE (Multimedia Interactive Learning Environment), was used for this purpose. MILE provides an advanced database to store all video episodes and interviews. In addition, it was also possible to store scans of student products and educational materials in an organized way in this database.

Table 2.4 Elements in the video portfolio

	Information sources	**Details**	**Aspect to be covered**
Primary sources	Video episodes	- Film fragments on the job while teachers held a coaching conference with a group of students	Professional performance
	Summary of each video episode	- What learning activity is coached by the teacher during the video episode - How far students are in completing the complex tasks - Summary of what happens during the video episode	Context information
	Task	- Description of the kind of task students work on during the video episode	Context information
	Interview with teacher	- What was the reason for supporting the students in …. ? - What did you aim to accomplish with the students? - In what way did you aim to accomplish …. ? - Why did you choose this approach? - Are you satisfied with the way you handled this situation? Why (not)?	Decisions underlying professional performance

Table 2.4 Elements in the video portfolio (Continued)

	Information sources	**Details**	**Aspect to be covered**
Primary sources	Interview with student(s)	- Did sir/madam …. help you to go on with ….. ? - In what way did/didn't he/she help you? - Do you think he/she helped you just in time with…. or do you think he/she could have helped you earlier or later with… ? Why do you think this? - Does sir/madam …. always help you in this way, or does he/she usually use a different approach? Can you give an example of a different approach used by sir/madam ….?	Consequences of teachers' actions
	Students' background information	Individual students: - Age - Current grade level - Unit of education - Previous training - Details of school career - Details of special needs Group of students: - Information on whether the students had worked together before - Reasons for putting these particular students together in one group	Context information
Secondary sources	Additional educational materials	- Information about how to organize meetings - Information about how to make minutes - Information about what should be included in proper planning	Context information
	Students' products	- Floor plans - Time schedules - Minutes	Context information

2.2.4 Scoring method for assessing video portfolios

Predominantly an analytical scoring method was used in this project. In an analytical approach, assessors start by scoring specific aspects of performance according to guidelines and criteria. Assessors then use the scores on specific aspects of the performance to build a judgment of the overall performance. In the scoring method

constructed, for example, assessors looked for evidence of constructive coaching and practice-oriented coaching in individual video episodes, and assigned a score to the entire performance in the video episode based on the evidence found. Furthermore, assessors built an overall judgment of teachers' coaching performance based on the scores for individual video episodes. Because analytic scoring methods are based on scoring guidelines and criteria, it is supposed that there is little room for assessors' personal views, beliefs, and opinions, and that it should lead to more objective and reliable judgments (Klein & Stecher, 1998). Guidelines for collecting evidence and criteria for evaluating the evidence collected were derived from the interpretive framework. Assessors were asked to score a video portfolio in four steps, as described in Table 2.5. During the scoring of the video portfolios, they used two different kind of score forms that are presented in Appendix 2 and 3.

The analytic scoring method was elaborated using the guidelines for a valid interpretation process introduced by Moss, Schutz, and Collins (1998) and Schutz and Moss (2004). The first guideline is that assessors should use all available evidence to base a judgment on. In accordance with this guideline, in steps two and three of the scoring method assessors are urged in advance to consider all available evidence and to check afterwards whether they based the assigned score on all available evidence. The second guideline is that assessors should actively search and consider counterevidence. In order to conform to this guideline, in step 1 of the scoring method assessors are urged to search for coaching interventions demonstrated by the teacher that do provide opportunities for students as well as interventions that do not. The third guideline assumes that valid interpretations derive from discussions with other assessors. In the discussions, assessors should challenge one another's interpretations, so that the acceptability and tenability of the interpretations are critically checked. In that way, the impact of selective observation, personal points of view, beliefs, and opinions should be reduced as much as possible. Based on this guideline, a fourth step was included in the scoring method in which assessors compared and discussed the scores assigned and the evidence and arguments on which the scores were based. After the consultation, the assessors could either hold on to their judgment(s) or make adjustments.

Table 2.5 Scoring method for judging video portfolios

Step 1 Collecting evidence from a video episode
Examine the following information sources in the video portfolio:
- Teachers' background information;
- Students' background information;
- Summary of the video episodes;
- Interview with the teacher.

Watch the video episode and answer the following questions:
- Which coaching interventions do or do not provide opportunities to improve students' performance of learning activities?
- Which coaching interventions do or do not provide opportunities for students to improve in constructing realistic perceptions of professional thinking and acting in practice?
- As the questions indicate, look for positive as well as negative evidence. Negative evidence pertains to coaching interventions that do not contribute to students' undertaking of learning activities and perceptions of professional thinking and acting in practice and/or missed opportunities in coaching.
- Take notes on the score form.
- Determine what interventions could be marked as (counter-) evidence for constructive and practice-oriented coaching.

Step 2 Assigning scores to teacher performance in a video episode
Consider all the available evidence for constructive as well as for practice-oriented coaching:
- What evidence is important, and what is less important?
- How can positive and negative evidence be counterbalanced?
- Does all evidence direct to a specific level of competence, or are contradictions perceived in the evidence?
- After you have assigned a score, check whether it represents all the available evidence.
- Assign a score to the coaching performance in the video episode, based on the performance levels for constructive and practice-oriented coaching.
- Write a brief summary in which you substantiate the scores assigned. In the summary, refer to or cite important arguments and evidence.

Step 3 Assigning an overall score to teacher performance across video episodes
- Assign an overall score for constructive and practice-oriented coaching based on the performance levels, for all video episodes concerning coaching aimed at a specific learning activity.
- The assigned overall score does not have to be equal to the average of all scores assigned to the individual video episodes, since you can weigh scores in order to correct for differences in video episodes with regard to complexity, or for differences in (extremely) high or low contributions to improvement in learning activities and perceptions of professional thinking and acting.
- In what way can the performance in the individual video episodes be counterbalanced?
- Does the entire performance direct to a specific level of competence, or are contradictions perceived?
- After you have assigned a score, check whether the score represents all the available evidence.
- Write a brief summary in which you comment on the scores assigned. In the summary, refer to or cite important arguments and evidence concerning individual video episodes.

Table 2.5 Scoring method for judging video portfolios (Continued)

Step 4 Consulting a fellow-assessor - After judging the video portfolios individually, discuss the assigned scores and written rationales with a fellow-assessor. - Compare assigned scores and explicitly discuss differences in assigned scores and cited evidence and arguments. - After the consultation, determine whether to stand by the original judgment(s) or to make adjustments.

2.2.5 Assessor training

Assessor training has emerged as a useful approach to promote more accurate ratings in performance assessments. Therefore, an assessor training course was set up to prepare assessors for scoring and judging video portfolios. Four training sessions were developed, aimed at enabling systematic and consistent use of the scoring method designed (design principle 7). Assessors were trained in each of the four steps of scoring a video portfolio and in applying the constructs from the interpretive framework.

Depending on the type of scoring and rating to be used, different kinds of assessor training have proven to be successful (Day & Sulsky; 1995; Stamoulis & Hauenstein, 1993). In the scoring procedure, it is important that assessors have common conceptualizations of what constitutes competent coaching, and that they are able to categorize performances into the same performance levels. In order to promote accuracy in categorizing performances, elements of Frame-of-Reference training were incorporated in the assessor training (Woerh & Huttcuff, 1994). Elements of Rating-Error-Training were also included in the training course to obtain awareness of rating errors and to avoid occurrence of these errors (Woerh & Huttcuff, 1994).

During the assessor training, video episodes that were not included in the video portfolios were observed and discussed. The scoring method was practiced step by step, and assessors received feedback about the following aspects:

- identifying, selecting, and quoting evidence from video episodes which is/is not consistent with the conceptual framework;
- evaluating evidence and reasoning about evidence in terms which are/are not consistent with the conceptual framework;

- assigning scores to video episodes which are/are not based on the designed performance levels for constructive and practice-oriented coaching (see Tables 2.2 and 2.3);
- evaluating performance across video episodes and reasoning about performance across video episodes in terms that are/are not consistent with the conceptual framework;
- assigning scores to the complete video portfolio which are/are not consistent with the conceptual framework.
- writing a rationale in which assigned scores are legitimized.

During the training course, much time was spent on discussing how to weigh evidence before assigning a score to a single video episode, and how to weigh performance across different video episodes before assigning an overall score.

2.3 Evaluation of the practical utility of the assessment procedure

2.3.1 Participants

Six assessors were selected who participated in the educational innovation MTS+ project and had experience in coaching students. These assessors were trained in scoring video portfolios as described in the previous section.

2.3.2 Procedure

The trained assessors scored the video portfolios designed as described in section two. Each assessor scored three of the four video portfolios, because scoring of all the portfolios would have taken too much time. The assessors installed the MILE software, including the video portfolios on their own computers, and scored the video portfolios independently and at their own pace. After scoring the video portfolios individually, they discussed the scored portfolios in pairs.

2.3.3 Instruments

In order to determine to what extent assessors agreed on the assigned scores to video portfolios based on the designed scoring method, filled out score forms were collected. For each video episode, two scores were assigned: one for constructive coaching and one for practice-oriented coaching. Furthermore, the assessors assigned

scores for constructive coaching and for practice-oriented coaching across video episodes concerning coaching aimed at cognitive learning activities; coaching aimed at meta-cognitive learning activities; coaching aimed at affective learning activities; and coaching aimed at learning activities with regard to collaborative learning.

To obtain more detailed information about factors that stimulated or hindered the assessors in making valid interpretations and judgments, they were interviewed. After scoring the three video portfolios, all assessors participated in a semi-structured interview about their experiences in using the assessment procedure. In the interview, the assessors were asked about four themes: 1) the composition of the video portfolios, 2) interpreting and judging video episodes and video portfolios, 3) the criteria and performance levels used, and 4) the scoring method as offered.

2.3.4 Analysis

A Gower coefficient was used as an estimate of interrater agreement for this discrete sample of assessors. A generalizability coefficient is usually used for this purpose. However, owing to the small variation in the assigned scores found in this study, a generalizability coefficient could not be used as an indicator of interrater agreement. The Gower coefficient is not sensitive to a lack of variance. The Gower coefficient is based on absolute differences between assigned scores (Zegers, 1989). The range of the Gower coefficient is from 0 (no agreement) to 1 (perfect agreement). Gower coefficients from 0.65 to 0.80 are perceived as an acceptable level of agreement. Gower coefficients lower than 0.65 represent low agreement, and Gower coefficients higher than 0.80 represent high agreement.

A content analysis was used to analyze the interview transcripts. Assessors' responses to the interview questions were searched for aspects that stimulated and hindered them in making interpretations and judgments for each theme. Issues raised by more than one assessor were summarized and exemplified using quotes.

2.4 Results

2.4.1 Interrater agreement

A Gower coefficient was determined for assigned scores, showing the extent to which the assessors assigned the same scores to constructive coaching and practice-oriented coaching in all video episodes (Table 2.6). A very high level of interrater agreement was found for assigned scores to practice-oriented coaching; because this type of coaching barely took place in the video episodes (or in practice), assessors consistently assigned the lowest score. The high levels of agreement with regard to practice-oriented coaching are, therefore, not representative and are not included in Tables 2.6, 2.7, and 2.8. Furthermore, the Gower coefficients were determined for scores assigned to constructive coaching in the video episodes across teachers (Table 2.7) and for overall scores assigned to constructive coaching across teachers (Table 2.8). The interrater agreement presented in Table 2.7 and 2.8 are based on three of the four teachers and four/five of the six assessors, because not all assessors scored all teachers due to the fact that scoring all teachers would have taken too much time.

Table 2.6 Gower coefficients for scores assigned to video episodes for individual teachers

	Constructive coaching
Teacher 1 (10 video episodes; 4 assessors)	0.67
Teacher 2 (10 video episodes; 4 assessors)	0.70
Teacher 3 (10 video episodes; 5 assessors)	0.73
Teacher 4 (8 video episodes; 4 assessors)	0.75

Table 2.7 Gower coefficients for scores assigned to video episodes across teachers

	Constructive coaching
Teachers 1, 3, and 4 (28 video episodes; 2 assessors)	0.67
Teachers 2, 3, and 4 (28 video episodes; 2 assessors)	0.73

The Gower coefficients presented in Table 2.6 show that an acceptable level of agreement was obtained for judging constructive coaching in individual video episodes. The results of the analysis across teachers (Table 2.7) support these results.

Table 2.8 Gower coefficients for overall scores across teachers

	Constructive coaching
Teachers 1, 3, and 4 (8 overall scores; 2 assessors)	0.81
Teachers 2, 3, and 4 (8 overall scores; 2 assessors)	0.96

The Gower coefficients presented in Table 2.8 show that a high level of assessor agreement was obtained for overall scores for constructive coaching. The results show that although assessors sometimes varied in their judgments of performance in specific video episodes, they agreed on teachers' performance across different video episodes.

2.4.2 Interview study

The results of the interview study are presented below according to the four themes addressed in the interview.

The composition of video portfolios

The assessors used for the most part the video episodes, interviews, summaries of the video episodes, and students' background information in scoring and judging the video portfolios. All assessors reported that, besides the video episodes, they considered the interviews as the most relevant source of evidence in the video portfolio. The assessors considered the interview with the teacher and the student(s) indispensable background information for judging the video episodes. The interview with the teacher was used mainly to retrieve information about what the teacher aimed to accomplish during the video episode. The assessors reported that this information

helped in directing observations to relevant aspects of performance and relevant consequences of teachers' performance. All assessors found the interview with the student(s) even more important. They indicated that especially this source of evidence provided instant proof for positive or negative consequences of teachers' actions. However, some assessors noted that not all students had been interviewed, so they could not determine whether the coaching had been effective or ineffective for all students. Furthermore, some assessors suspected that students had given socially acceptable answers, which would have compromised the evidence.
Most assessors also indicated that the brief summaries of the content of the video episodes provided useful information, as it helped in directing attention to relevant evidence. One assessor reported: "That summary works well, because then you know what is going to happen and you can work through the descriptions of relevant learning activities and the examples of coaching interventions before you watch the video episode. Then you have it all in your head and you know what to look for."

Interpreting and judging video episodes and video portfolios
Assessors found it hard to evaluate teachers' contributions to positive consequences for students based on single video episodes. One assessor reported: "For some video episodes it is hard to evaluate teachers' contributions. Sometimes I would have liked to see the students in the future, how they handled a comparable situation in the future, to see whether they had improved or not. You just see a bit of what happens. It was only in the portfolio of teacher 3 that you could see a certain development during the video episodes. In that portfolio you could follow students' development."

Assessors indicated that especially the first video episodes of a video portfolio were hard to assess. They reported that it was especially difficult to identify evidence in the beginning. They indicated that they used the descriptions of the learning activities and the examples of coaching interventions a lot, in order to keep in mind what to look for. They felt that as they evaluated more situations, they became more skilful. Furthermore, they reported that the first few video episodes of a portfolio were hard to evaluate, because they did not yet have a point of reference. One assessor stated: "For those episodes, I have to guess and assume. It is the first situation I have seen, after all. The more video episodes I observed from a specific teacher, the more familiar I got with his or her method."

It seems that some video episodes are easier to score and judge than others; 'straightforward' coaching video episodes are easier to score. One assessor stated: "[…] It depends on what video episode you have to evaluate. Some episodes are clear, less complex. Then it is easier to fill in scorecard 1." Assessors noted that some factors made coaching situations more straightforward and, therefore, easier to score and judge; the first factor they indicated, was when teachers' behavior in the video episodes matched teachers' intentions explained in the interview. One assessor stated: "When teachers' behavior and reported intentions match, you can understand what the teacher aims to do in the coaching situation, which makes it easier to score". Another factor indicated was when coaching in a specific learning activity could clearly be distinguished from coaching in another learning activity. A third factor indicated by assessors was that coaching situations in which students needed support only in one specific learning activity were easier to score.

Assessors indicated that video episodes of five to ten minutes provided enough information on teachers' performance in that situation. One assessor reported: "I noticed that during video episodes that were longer than ten minutes, my attention lapsed and I no longer noticed all the evidence. In ten minutes, I saw enough evidence to form a judgment on anyway."
Assessors rarely watched a whole video episode more than once, but most assessors watched some parts of a video episode a second time. They indicated that it was too time consuming to watch a video episode a second time, but they viewed parts of it for a second time in order to check what really happened and whether they overlooked evidence or not.

Most assessors indicated that, after assessing six video episodes, they had developed a clear view on the teachers' coaching competence. One assessor reported: "After viewing a video episode you get familiar with the approach that the teacher uses in coaching students, and after five or six video episodes you have seen enough to base a score on."

Positive evidence for practice-oriented coaching was hardly found in video episodes. Assessors stated that this type of coaching was scarcely to be found in the video episodes, so they assigned level 1 to almost all video episodes. One assessor indicated that he sometimes wondered whether it was fair to assign the lowest score based only on negative evidence in terms of missed opportunities. He said: "Sometimes I

thought, is it fair to assign level 1 when practice-oriented coaching doesn't take place? And do you have to perform this type of coaching in all coaching situations?"

The criteria and performance levels

Most assessors indicated that the performance levels were useful, but some expressed the view that the assignment of scores remains speculative. Most assessors indicated that the descriptions of the performance levels were useful in assigning scores. One assessor reported that it helped him in being objective. He stated: "I already had certain ideas about the teachers in the portfolios, but by judging performance based on the performance levels, I managed to block out some of these biases."

Some assessors indicated that the difference between performance on level 2 and performance on level 3 was hard to distinguish. In most cases, extreme performances (levels 1 and 4) were easy to recognize, especially coaching on level 1. Some assessors indicated that it was sometimes also hard to distinguish level 3 from level 4.

The scoring method

The assessors considered the scoring method to be complex and time consuming, but indicated that as they judged more video episodes and portfolios, they became more proficient in it. Assessors needed approximately 18 hours to assess three portfolios. Some assessors indicated that it was essential to practice using the scoring method, especially collecting evidence and applying the assessment scales to constructive and practice-oriented coaching. One of them stated: "You can't assess a video portfolio without training; it is too complex."

2.5 Conclusion and discussion

The aim of the study was to investigate how well assessors were able to cope with the assessment procedure designed, and to explore how they were supported by this procedure in making valid interpretations and judgments based on video portfolios. In order to answer our research questions, the interrater agreement was determined for scores assigned to video episodes and for overall scores, and assessors were interviewed about their experiences of scoring and judging video portfolios.

Based on the acceptable and high level of interrater agreement found for assigned scores to video episodes and assigned overall scores, it seems that the assessors were reasonably capable of using the assessment procedure. To arrive at these levels of agreement, assessors needed substantial training in using the assessment procedure, which after all took a lot of energy. Particularly recognizing evidence in the video portfolio and getting familiar with the steps in the scoring method took time.

The results of the interview study provided more detailed information about the practical utility of the video portfolios. The assessors mentioned three factors that assisted them in making valid interpretations and judgments. First, the descriptions of learning activities and related coaching interventions helped assessors in identifying relevant coaching interventions in the coaching performance, especially in the beginning. Second, the summaries of what happened during the video episodes seemed to help assessors in directing their attention to relevant aspects of teachers' coaching performance. In the light of theories with regard to the use of schemata by assessors while they assess (DeNisi, Cafferty, & Meglino, 1984; Feldman, 1981; Landy & Farr, 1980), the findings indicate that the descriptions, examples, and summaries might have activated relevant schemata and constructs in the assessors' minds, and might have assisted them in applying these constructs during the scoring of the video portfolios. A second factor assessors mentioned that helped them in making valid interpretations and judgments was the information added to the video episodes. Particularly the interviews with the teachers, which informed assessors about the decisions underlying the performance, and the interviews with the students, which informed assessors about the impact of teachers' actions, were perceived as indispensable background information for making interpretations and judgments. These findings suggest that especially the information provided by the interviews was essential to assessors for understanding and interpreting the performance in the video episode (Heller et. al., 1998; Schutz & Moss, 2004). A third factor that helped assessors in making valid judgments pertains to the nature of the video episodes. Assessors noted that it was easier to make interpretations and judgments about straightforward coaching episodes. This finding is in line with the findings of Heller et al. (1998) and Schutz and Moss (2004), which show that it is hard for assessors to develop a coherent representation of a portfolio with inconsistent or ambiguous evidence. This is a difficult problem to address; inconsistent or ambiguous portfolios should also be judged. Special measures can be taken, however, when assessors indicate that a video portfolio or episode is 'inconsistent' or 'ambiguous'. For example,

these portfolios can be judged by a larger committee of assessors, or more video episodes or more context information, or both, can be included (Schutz & Moss, 2004).

Some disabling factors were also mentioned. First, the assessors considered single video episodes hard to assess. They claimed that the single video episodes represented just a part of what happened. When assessors observed five or six video episodes, they got a clear view of teachers' coaching as long as a certain degree of variety in the video episodes was established. This finding can be explained by the theory introduced by Schutz and Moss (2004), according to which assessors search for a pattern in the data. It seems that the evaluation of a single video episode leaves too many blank spots to allow an assessor to discover a pattern in teachers' coaching competence. Five to six episodes, on the other hand, seem to provide assessors with enough data to build a coherence pattern. However, the claim that five or six episodes should be enough for making valid interpretations and judgments is merely an indication made by assessors. This claim should be verified in future research using quantitative analyses. Second, video episodes lasting longer than 15 minutes do not seem to contribute to more valid interpretations and judgments. The assessors reported that this was mainly because it was hard to concentrate for longer than 15 minutes, and that no new information about teachers' coaching was added during the rest of the video episode. This finding is in line with the literature on concentration span during lectures (Bligh, 1979). Research has shown that students' concentration span during a lecture slowly decreases. After 20 minutes students' attention had dropped to 50%. Based on these results and our findings, it seems that 10 to 15 minutes should be a maximum length for video episodes in video portfolios. Third, assessors found it difficult to distinguish coaching on score level 2 from that on score level 3. This was the critical distinction between a 'negative' and a 'positive' judgment in the assessment procedure designed. During the training course, much time was spent discussing what performance level was appropriate to assign to video episodes; this research finding indicates that more attention should have been given to the differences in performance between score levels 2 and 3. A fourth factor that hindered assessors in making interpretations and judgments was that a disproportionate amount of negative evidence in terms of missed opportunities was provided for practice-oriented coaching. This finding illustrates that it is wise to conduct a job analysis before constructing a video portfolio, in order to explore which situations elicit performance that holds evidence for the

domain of competence to be assessed (Mislevy, Steinberg, Breyer, Almond, & Johnson, 2002). It could be that practice-oriented coaching was not taken place, because they were all beginning coaches, with between one and two years of experience in coaching students. It is likely that beginning coaches first focus on mastering constructive coaching. They then switch their attention to practice-oriented coaching. The participants in this research were probably still at a point where they were so occupied with constructive coaching that they were not able to pay attention to practice-oriented coaching. The results presented in this study showed that practice-oriented coaching could not be scored in a proper way based on the video portfolios. Therefore, the scoring of practice-oriented coaching was left out of the scoring procedure in the studies that are presented in chapter three and four of this dissertation.

Future research

The practical utility of video portfolios was examined in this study. The rater agreement (n=6) was determined for scores assigned to video episodes and overall scores, and an overview was presented of aspects that stimulated or hindered assessors in making valid interpretations and judgments. In order to obtain a complete view of the quality of the performance assessment, further investigation of the reliability and validity of the assessment procedure is essential. To acquire more robust indications of the reliability of the assessment procedure, supplementary quantitative analyses are needed based on a larger sample of assessors. Furthermore, to investigate the validity of the assessment procedure, additional qualitative analyses are needed of the evidence and arguments assessors use to legitimize the scores they assign. Based on the findings of these analyses, it can be determined whether the descriptions, examples, and summaries really contribute to valid interpretations and judgments. It can also be examined on the basis of these findings whether and in what way the nature of video episodes affects the validity of interpretations and judgments of the video episodes.

References

Barton, J., & Collins, A. (1993). Portfolios in teacher education. *Journal of Teacher Education*, *44*, 200-210.

Barton, J., & Collins, A. (1997). *Portfolio assessment: A handbook for educators.* Dale Seymour Publications.

Beijaard, D. & Verloop, N. (1996). Assessing teachers' practical knowledge. *Studies in Educational Evaluation, 22*, 275-286.

Boekaerts, M. (1999). Self-regulated learning: Where we are today. *International Journal of Educational Research*, *31*, 445-457.

Boekaerts, M., & Simons, P.R.J. (1995). *Leren en instructie. Psychologie van de leerling en het leerproces (Learning and instruction. Psychology concerning student and students' learning process).* Tweede druk. Assen: Van Gorcum.

Bolhuis, S. (2000). *Naar zelfstandig leren: Wat doen en denken docenten (Towards self-regulated learning: What teachers do and think).* Apeldoorn: Garant.

Bligh, D.A. (1979). *What's the use of lectures?* Harmondsworth.

Brown, A.L., & Campione, J.C. (1994). Guided discovery in a community of learners. In K. McGilly (Ed.), *Classroom lessons: Integrating cognitive theory and classroom practice* (pp. 229-270). Cambridge, MA: MIT Press/Bradford Books.

Butler, D.L., & Winne, P.H. (1995). Feedback and self-regulated learning: A theoretical syntheses. *Review of Educational Research*, *65*(3), 245-281.

Collins, A., Brown, J.S., & Newman, S.E. (1989). Cognitive apprenticeship: Teaching the crafts of reading, writing, and mathematics. In L.B. Resnick (Ed.), *Knowing, learning and instruction: Essays in honor of Robert Glaser* (453-494). Hillsdale, NJ: Erlbaum.

Darling-Hammond, L., & Snyder, J. (2000). Authentic assessment of teaching in context. *Teacher and Teacher Education, 16,* 523-545.

Day, D.V., & Sulsky, L.M. (1995). Effects of frame-of-reference training and information configuration on memory organization and rating accuracy. *Journal of Applied Psychology, 80*, 158-167.

DeNisi, A.S., Cafferty, T.P., & Meglino, B.M. (1984). A cognitive view of the performance appraisal process: A model and research propositions. *Organizational Behavior and Human Performance, 33*, 360-396.

Duffy, T. M., Lowyck, J., & Jonassen, D. H. (Eds.) (1993). *Designing environments for constructive learning.* New York: Springer Verlag.

Dwyer, C.A. (1993). Teaching and diversity: Meeting the challenges for innovative teacher assessments. *Journal of Teacher Education, 44*(2), 119-129.

Dwyer, C.A. (1994). Criteria for performance-based teacher assessment: Validity, standards, and issues. *Journal of Personnel Evaluation, 8*, 135-150.

Dwyer, C.A. (1998). Psychometrics of Praxis III: Classroom performance assessments. *Journal of Personnel Evaluation in Education, 12*(2), 163-187.

Feldman, J.M. (1981). Beyond attribution theory: Cognitive processes in performance appraisal. *Journal of Applied Psychology, 66*, 127-148.

Frederiksen, J.R., Sipusic, M., Sherin, M., & Wolfe, E.W. (1998). Video portfolio assessment of teaching. *Educational Assessment*, *5*(4), 225-298.

Gipps, C.V. (1994). *Beyond testing: Towards a theory of educational assessment*. London, Washington D.C.: The Falmer Press.

Girod, G. (Ed.) (2002). *Connecting teaching and learning. A handbook for teacher educators on teacher work sample methodology*. Monmouth: Western Oregon University, Washington DC: American Association of colleges for teachers. ERIC Clearinghouse on Teaching and Teacher Education, ED 463 282.

Haertel, E.H. (1991). New forms of teacher assessment. *Review of Research in Education, 17*, 3-19.

Heller, J.I., Sheingold, K., & Myford, C.M. (1998). Reasoning about evidence in portfolios: Cognitive foundations for valid and reliable assessment. *Educational Measurement, 5*(1), 5-40.

Johnson, D., & Johnson, R. (1994). *Learning together and alone: cooperative, competitive, and individualistic learning* (4th ed.). Boston: Allyn & Bacon.

Kagan, D.M. (1990). Ways of evaluating teacher cognition: Inferences concerning the Goldilocks principle. *Review of Educational Research, 60*, 419-469.

Kane, M. (1992). An argument-based approach to validity. *Psychological Bulletin, 112*, 527-535.

Kelly, G.A. (1995). *The psychology of personal constructs*. New York: Norton.

Klein, S.P., & Stecher, B.M. (1998). Analytic versus holistic scoring of science performance tasks. *Applied Measurement in Education, 11*(2), 121-137.

Landy, F.J., & Farr, J.L. (1980). Performance rating. *Psychological Bulletin, 87*, 72-107.

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

Lave, J., & Wenger, E. (1991). *Situated Learning: Legitimate peripheral participation*. Cambridge: Cambridge University Press.

Mislevy, R.J., Steinberg, L.S., Breyer, F.J., Almond, R.G., Johnson, L. (2002). Making sense of data from complex assessments. *Applied Measurement in Education, 15*(4), 363-389.

Moerkamp, T., De Bruijn, E., Van der Kuip, I., Onstenk, J., Voncken, E. (2000). *Krachtige leeromgevingen in het MBO. Vernieuwingen van het onderwijs in beroepsopleidingen op niveau 3 en 4 (Powerful learning invironments in senior secondary vocational education. Educational innovations in vocational education on level 3and 4)*. Amsterdam: SCO-Kohnstamm Instituut.

Moss, P.A. (1994). Can there be validity without reliability? *Educational Researcher*, *23,* 5-12.

Moss, P.A., Schutz, A.M., & Collins, K.A. (1998). An intergrative approach to portfolio evaluation for teacher licensure. *Journal of Personnel Evaluation in Education, 12*(2), 139-161.

Onstenk, J. (2000). *Op zoek naar een krachtige beroepsgerichte leeromgeving: Fundamenten voor een onderwijsconcept voor de bve-sector (A search for powerful learning environments: A basis for a teaching philosophy in senior secondary vocational education).* 's-Hertogenbosch: CINOP.

Perry, N., Phillips, L., & Dowler, J. (2004). Examining features of tasks and their potential to promote self-regulated learning. *Teachers College Record, 106*, 1854-1878.

Perry, N.E. (1998). Young children's self-regulated learning and the context that support it. *Journal of Educational Psychology, 90*, 715-729.

Roelofs, E., & Sanders, P. (2007). Towards a framework for assessing teacher competence. *European Journal for Vocational Training*, *40*(1), 123-139.

Salzman, S.A., Denner, P.R., Bangert, A.W., & Harris, L.B. (2001). *Connecting teacher performance to the learning of all students: Ethical dimensions of shared responsibility.* Pocatello, Idaho: Idaho State University; ERIC Reproduction services ED 451182.

Schaaf, van der, M.F., Stokking, K.M., & Verloop, N. (2005). Cognitive representations in raters' assessment of teacher portfolios. *Studies in Educational Evaluation, 31*, 27-55.

Schalock, H.D., Schalock, M., & Girod, G. (1997). Teacher work sample methodology as used at Western Oregon University. In J. Millman (Ed.)., *Grading teachers, grading schools: Is student achievement a valid evaluation measure?* (15-45). Newbury Park, CA: Corwin Press.

Schutz, A.M., & Moss, P.A. (2004). Reasonable decisions in portfolio assessment: Evaluating complex evidence of teaching. *Education policy Analysis Archives, 12*(33). Retrieved 7/19/2004 from http://epaa.asu.edu/v12n33/.

Seldin P. (1991). *The teaching portfolio.* Bolton: MA. Anker Publishing Company, Inc.

Shuell, T. J. (1993). Toward an integrated theory of teaching and learning. *Educational Psychologist*, *28*, 291–311.

Shulman, Lee (1998). Teacher portfolios: A theoretical activity. In N. Lyons (Ed.), *With portfolio in hand.* (pp. 23-37) New York: Teachers College Press.

Slavin, R. (1990). *Cooperative learning: theory, research, and practice.* Englewood Cliffs: NJ, Prentice-Hall.

Stamoulis D.T., & Hauenstein, N.M.A. (1993). Rater training and rater accuracy: Training for dimensional accuracy versus training for rater differentiation. *Journal of applied psychology, 78*(6), 994-1003.

Uhlenbeck, A.M. (2002). *The development of an assessment procedure for beginning teachers of English as a foreign language.* Doctoral dissertation. Leiden: ICLON Graduate School of Education.

Uhlenbeck A.M., Verloop N., & Beijaard D. (2002). Requirements for an assessment procedure for beginning teachers: Implications from recent theories on teaching and assessment. *Teachers College Record, 104*(2), 242-272.

Vermunt, J.D., & Verloop, N. (1999). Congruence and friction between learning and teaching. *Learning and Instruction, 9,* 257-280.

Vermunt, J., & Verschaffel. (2000). Process oriented teaching. In P.R.J. Simons, J. van der Linden & T. Duffy. *New Learning* (pp. 209-225). Dordrecht: Kluwer Academic Publishers.

Vygotsky, L.S. (1978). *Mind in society: The development of higher psychological processes*. Cambridge, MA: Harvard University press.

Wade, R. C., & Yarbrough, D. B. (1996). Portfolios: A tool for reflective thinking in teacher education? *Teaching and teacher education*, *12*, 63-79.

Winne, P.H., & Hadwin, A.F. (1998). Studying as self-regulated learning. In D.J. Hacker, J. Dunlosky, & A.C. Graesser (Eds.), *Metacognition in educational theory and practice* (pp. 277-304). Hillsdale, NJ: Erlbaum.

Woerh, D.J., & Huffcutt, A.I. (1994). Rater training for performance appraisal: A quantitative review. *Journal of Occupational and Organizational Psychology, 64*, 189-205.

Zegers, F.E. (1989). Het meten van overeenstemming (Measuring interrater agreement). *Nederlands Tijdschrift voor de Psychologie, 44*, 145-156.