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

Bakker, M.E.J.

Design and evaluation of video portfolios

Reliability, generalizability, and validity of an authentic performance assessment for teachers

ICLON

Leiden University Graduate School of Teaching

ico

This research was carried out in the context of the Interuniversity Center for Educational Research

NWO

Netherlands Organisation for Scientific Research

This research was funded by the Netherlands Organization for Scientific Research (NWO) (Project no. 411-02-207).

Title: Design and evaluation of video portfolios: Reliability, generalizability, and validity of an authentic performance assessment for teachers
Titel: Design en evaluatie van videodossiers: De betrouwbaarheid, generaliseerbaarheid en validiteit van competentiebeoordelingen bij docenten

Print: Mostert & Van Onderen! Leiden
Cover design: Simon Ouwerkerk (www.ouwerkerk.tv)
ISBN 978-90-804722-8-0

Design and evaluation of video portfolios

Reliability, generalizability, and validity of an authentic performance assessment for teachers

Proefschrift

ter verkrijging van

de graad van Doctor aan de Universiteit Leiden,

op gezag van Rector Magnificus prof. mr. P.F. van der Heijden,

volgens besluit van het College voor Promoties

te verdedigen op dinsdag 2 december 2008

klokke 13.45 uur

door

Maria Elisabeth Jacomina Bakker

geboren te Beverwijk

in 1977

Promotiecommissie

Promotores
Prof. dr. N. Verloop
Prof. dr. D. Beijaard

Copromotores
Dr. E.C. Roelofs
Dr. E.H. Tigelaar

Referent
Prof. dr. J.M.G. Brekelmans

Overige leden
Prof. dr. P.F. Sanders
Prof. dr. J.H. van Driel
Prof. dr. H. Hulshof

Table of contents

Chapter 1

Introduction

The research study presented in this dissertation focuses on issues pertaining to the reliability, generalizability, and validity of authentic performance assessments of teachers. Generally, these assessments include the use of complex and open-ended assessment tasks, often carried out in varying contexts. Compared with more 'traditional' forms of assessment, in which the assessment tasks used are more distant from the classroom reality, the context and tasks present in authentic performance assessments bring specific threats to the methodological quality of these assessments. Central to this dissertation are three studies addressing processes and factors that affect the reliability, generalizability, and validity of authentic performance assessments. The focus is particularly on issues related to the process of construction of performance assessments, assessors' scoring, and the reliability, generalizability, and validity of performance scores. The three studies are based on data that was collected during and after assessors' scoring of a structured video portfolio. Video portfolios were developed for assessing teachers' coaching competence while coaching students in senior secondary vocational education. As part of this assessment procedure, trained assessors judged the video portfolios of four teachers. A structured video portfolio consists of a collection of evidence pertaining to the competence to be assessed, i.e., teachers' coaching competence while coaching students in senior secondary vocational education. The main source of evidence consisted of video recordings that were systematically selected and showed teachers' coaching performance in key situations in practice. Additional sources of information were included in the portfolio pertaining to the context in which the coaching took place, like information about how far students were in completing an assignment, information about students' backgrounds, or an interview with the teacher about the decisions underlying his or her actions. This first chapter depicts the theoretical and practical background to this dissertation, the general and specific research questions, the context, and the relevance of the studies. The chapter concludes with an overview of the subsequent chapters.

1.1 Background to the study

1.1.1 New forms of teacher assessment

New forms of teacher assessment have been developed in recent decades in response to dissatisfaction with both content and formats of existing assessments. These new forms of assessment are often referred to as ‘authentic performance assessments’ (Darling-Hammond & Snyder, 2000; Haertel, 1991). The most important characteristic of performance assessments is that they are grounded in a more professional model of teaching in which the complexity of teaching and teachers’ context-specific decisions and actions are acknowledged. Teaching is viewed as complex because it involves immediate and adequate decision-making and acting that fit in the specific situation. It is also recognized that teaching 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 (Darling-Hammond & Snyder, 2000; Gipps, 1994; Haertel, 1991). Furthermore, teaching is viewed not only in terms of demonstrating adequate behavior or applying relevant knowledge, but as an integrated concept denoting teachers’ knowledge, skills, and attitudes that the teacher calls upon while performing in a specific context (Eraut, 1994; Gonzi, 1994). These conceptions of teaching have implications for the design of assessment procedures used to assess teaching. First, instead of assessing separate components of teaching, the use of multiple methods is recommended to cover the different aspects of teaching, such as knowledge, decisions, and actions (Uhlenbeck, 2002). Second, assessment tasks need to reflect the complexity of teaching. This requires that open-ended assessment tasks should be used that elicit complex decision-making and acting (Darling-Hammond & Snyder, 2000; Dwyer, 1998; Gipps, 1994; Haertel, 1991; Uhlenbeck, 2002). Third, the assessment should take place in a context that resembles the actual teaching context (Darling-Hammond & Snyder, 2000; Dwyer, 1998; Gipps, 1994; Haertel, 1991; Uhlenbeck, 2002).

Along with changing conceptions of teaching, new methods for authentic teacher performance assessment have been developed, which use direct evidence collected in realistic job situations. Typically, in a performance assessment, the teacher will be asked to perform, produce, or create something over a sufficient duration of time to permit evaluation of either the process or the product of performance, or both (Messick, 1994). In these new methods for authentic performance assessments, the focus is on collecting and judging evidence connected with the same sample of teaching situations. The evidence can pertain to teachers' actions, teachers' decision making, and students' learning. The use of connected evidence is visible in the methods of 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, particularly, video portfolios (Frederiksen, Sipusic, Sherin, & Wolfe, 1998).

1.1.2 Validity and reliability issues in authentic performance assessment

The new forms of assessment bring along new problems concerning validity and reliability. An important difference compared to traditional assessments is that in performance assessments assessors are used who judge the performance shown by the respondents during the assessment. As a result of the inclusion of human judgment in the assessment, specific pitfalls and threats arise. Assessors have a personal history, beliefs, and opinions which may affect their judgments of performance. In judging performance, it appears to be difficult for assessors to exclude biases stemming from their personal backgrounds and to prevent selective observation. Furthermore, in the new forms of assessment, open-ended and complex tasks are used that are situated in varying contexts. The respondents can react to those open-ended tasks in very different ways. It is not easy for assessors to score these performances in a consistent way (Gipps, 1994; Moss, 1994). These issues concerning the scoring of performance by assessors form a serious threat to the objectivity, reliability, and validity of the assessment outcomes. Another important difference from traditional assessment is the nature of the assessment tasks. Even when the assessment tasks come from the same domain, respondents show very divergent performances on these tasks. Issues pertaining to task specificity are probably the biggest threat to the validity of authentic performance assessments. So far, little is known about the actual causes of the variation in respondents' performance on different assessment tasks (Brennan, 2000; Dunbar, Koretz, & Hoover, 1991; Linn, 1994; Linn, Baker, & Dunbar, 1991; Linn & Burton, 1994; Miller & Linn, 2000; Ruiz-Primo, Baxter, & Shavelson, 1993; Shavelson, Baxter, & Gao, 1993). The selection of representative samples of assessment tasks is another important issue in authentic performance assessment. Complex and open-ended tasks are very time consuming. As a result, only a small number of tasks can be included in the performance assessment. Because of the small number of tasks, it is difficult to establish a representative sample of assessment tasks that covers all types of situations and relevant aspects of teaching that teachers face in practice (Brennan, 2000; Dunbar, Koretz, & Hoover, 1991; Linn, Baker, & Dunbar, 1991; Linn & Burton, 1994; Miller & Linn, 2000; Ruiz-Primo, Baxter, & Shavelson, 1993; Shavelson, Baxter, & Gao, 1993).

1.1.3 Validity arguments for authentic performance assessments

The most important consideration relating to the quality of assessment procedures pertains to validity. The primacy of validity is emphasized in professional standards and reaffirmed in most books and articles on assessment. Recent efforts to build a more coherent and unified view of validity have expanded its scope and further strengthened its importance (Kane 1992, 2004, 2006; Linn, 1994; Messick, 1989, 1994; Shepard, 1993). This breadth is evident in Messicks' definition of validity: "Validity is an integrated evaluative judgment of the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of inferences and actions based on test scores or other modes of assessment (1989, p. 13)." Although some objections to this definition can be raised, like that it is very broad (Borsboom & Mellenbergh, 2004), it is now generally accepted.

Messick described six aspects of construct validity that can be taken care of in a validation study: the content, substantive, structural, consequential, generalizability, and the external aspect. The content aspect of validity examines the content relevance and representativeness. Important in this aspect is that the assessment tasks elicit on all relevant aspects of a domain of competence and that the assessment covers the domain content. The substantive aspect refers to the theoretical rationales for the observed (in)consistencies in assessment responses, including process models for task performance, along with empirical evidence that the theoretical processes actually take place during performance of an assessment task. In this aspect, it is important that the thinking processes underlying the performance in the assessment tasks be comparable to the thinking processes underlying performance in practice. The structural aspect concerns the fidelity of the assessors' scoring procedures and scoring processes during the assessment. The consequential aspect is related to the consequences of the assessment for the person being assessed. The generalizability aspect examines the extent to which score properties or interpretations can be generalized to and across population, settings, and tasks. The external aspect refers to the relationship between scores obtained in the assessment and scores obtained in other assessments used to measure the same construct.

Although the Messick framework undoubtedly contains relevant aspects of validity, it poses the practically inclined assessment developer with difficulties in applying the framework to assessment programs. The framework does not offer practical guidelines as to how to arrive at conclusions during a validation study, as critics put it (Brennan, 1998; Crocker, 2003; Kane, 1992, 2004, 2006). In response to this problem, Kane proposed an eight-step argument-based approach to validity in which he offers some practical criteria for prioritizing different kinds of evidence that can be used to validate assessment procedures, including performance assessments. In his argument-based approach, Kane states that the validity of a performance assessment can be investigated by evaluating the chain of inferences that takes place when interpreting the outcomes of an assessment. Three inferences form the heart of the validity argument: (1) reliable and valid scoring of performance by assessors, (2) generalization from the observed score on an assessment task to a universe of assessment tasks, (3) extrapolation of assessment results to practice. In a thorough validity investigation, the tenability of all three inferences should be examined.

The first inference pertains to the scoring of the performance of respondents by the assessor: are the assessors' interpretations and judgments of the performance valid and reliable? As stated above, the influence of personal characteristics on judgments, such as selective observation, biases, and personal prejudices, can be a serious threat to the tenability of the first inference regarding scoring (Gipps, 1994; Moss, 1994). In determining the tenability of the second inference, the following question is relevant: does the score obtained on the basis of the assessment tasks represent the score that a respondent would have achieved if he or she had accomplished all possible tasks to measure the construct that is to be measured? In examining the third inference, the question is to what extent it is possible to extrapolate the performance measured in the assessment to performance outside the assessment context.

In the studies reported in this dissertation, the quality of a constructed performance assessment was evaluated by investigating the inferences distinguished by Kane. The quality of the video portfolio performance assessment was evaluated internally (Lissits, & Samuelson, 2007). This means in essence that the first two inferences of Kanes' chain of inferences were evaluated. Relations between assessment scores and external measures or criteria were not examined. Following Lissits and Samuelson (2007), evaluation of the internal validity of a performance assessment can be regarded as a critical initial activity in evaluating the quality of a performance assessment.

The research was started by designing an authentic performance assessment. In constructing an assessment, the design activities can be structured in such a way that validity evidence will emerge. Much of the work during this stage involves constructing representation in designing tasks and scoring procedures in addition to setting the boundaries of the performance domain. Thus, evidence for a valid performance assessment can be collected during the design of the performance assessment, in addition to evaluation of its validity afterwards. This process is referred to as a 'design-argument' for a valid performance assessment (Mislevy, 2007). In addition to the evaluation of validity based on Kane's chain of inferences, we show how the reliability, generalizability, and validity are warranted in the design of the performance assessment constructed.

1.1.4 Measures to reduce the threats to validity

In order to minimize the occurrence of threats to reliability, generalizability, and validity in authentic performance assessments, several measures have been proposed in the literature. The following (types of) measures are expected to have a positive effect on the validity of authentic performance assessments.

(1) Scoring guide based on a conceptual framework

A major measure pertains to the use of a scoring guide that includes appropriate criteria and performance levels. The scoring guide needs to be derived from a conceptual framework in which the construct to be assessed is defined. Important in constructing a scoring guide is that criteria describe essential aspects of the competence performance in terms of what professional teachers should know and be able to do. Performance levels should indicate to what extent teachers demonstrate knowledge and behavior that are defined in the criteria (Dwyer, 1993; Kagan, 1990). Furthermore, criteria and performance levels have to be formulated on an appropriate level of specificity. If the criteria and performance levels are formulated too broadly/generally, then it is difficult for assessors to apply these criteria and performance levels consistently. If they are formulated too narrowly/specifically, then there is the risk of getting lost in specifics so that the essence of teaching is missed. Another measure to establish appropriate criteria and performance levels is that the criteria and performance levels should be based on theoretical notions as well as on teaching in practice (Uhlenbeck, 2002). Furthermore, criteria and performance levels should specify what aspects of teaching are to be assessed, and not how teachers should carry them out (Darling-Hammond & Snyder, 2000; Dwyer, 1998; Gipps, 1994; Haertel, 1991; Uhlenbeck, 2002). This directs assessors to the core of teaching and not to the style of teaching. In this dissertation it was aimed to apply the above-mentioned measures by

- formulating criteria in terms of what the teacher should achieve;
- using relevant criteria formulated on an appropriate level of specificity;
- including only aspects that distinguish competent from less competent performance in the performance levels;
- adjusting the literature-based criteria and performance levels to practice by discussing them with teachers and through classroom observations.

(2) Assessor scoring

Regarding the scoring of performance, the following measures can be taken. First, use of a large number of assessors contributes to more reliable scoring (Kane, 2006). When multiple assessors judge a performance, the personal influence on the judgment of individual assessors decreases, so that the scores assigned are more accurate. Twelve assessors participated in this study. This relatively large number of assessors was used in order to be able to determine the minimum number of assessors needed for an acceptable level of interrater agreement. Second, to reduce the risk of invalid and unreliable judgments, it is advisable to use a systematic and transparent scoring procedure (Frederiksen & Collins, 1989; Linn, Baker, & Dunbar, 1991). In this study, this measure was applied by using a detailed scoring procedure that started with the scoring of specific aspects of the performance; a judgment was subsequently assigned to the whole performance based on guidelines and criteria. We elaborate on the scoring procedure in section 1.3.3.

(3) Assessor training

A factor that has a positive influence on the application of the criteria, performance levels, and scoring rules is assessor training (Day & Sulsky, 1995; Stamoulis & Hauenstein, 1993). In this dissertation, we aimed to apply this measure by setting up an assessor training course consisting of four training sessions. In section 1.3.4, we elaborate on this training.

(4) Standardizing the assessment task

A measure that is expected to overcome task specificity is the standardization of assessment tasks (Kane, 2006). As a result of standardizing assessment tasks, the openness of the tasks is reduced in order to direct the assessors to more homogeneous responses from respondents. This measure is expected to contribute to the reliability of the assessment outcomes. However, reducing the openness of the assessment tasks can directly harm validity of the assessment outcomes. It is a matter of finding a balance between open-ended tasks and standardized tasks (Kane, 2006). In the present study, video recordings of teachers' teaching in practice were judged by assessors. The video recordings were very authentic, but it was aimed to standardize them by selecting only video episodes showing a 'key situation', i.e., a situation that calls upon essential aspects of the competence to be measured in the performance assessment.

(5) Number of tasks included

A simple and highly effective measure to increase the representativeness of the sample of assessment tasks is to raise the number of tasks included in the assessment (Dunbar, Koretz, & Hoover, 1991; Ruiz-Primo, Baxter, & Shavelson, 1993). In practice, however, this measure cannot often be applied because of the time and costs involved. This remains a difficult issue in authentic performance assessment for which no clear-cut solution is at hand. In this study, the aim was to apply this measure by including ten video episodes in a video portfolio. Ten video episodes is a considerable number of video recordings in order to create variation in situations, but can be scored by assessors within a reasonable amount of time.

1.2 Problem definition and research questions

Much attention is currently given to the design, use, and methodological quality of performance assessments. As mentioned above, design principles ('measures') for realizing valid and reliable performance assessments are proposed in the literature. Furthermore, empirical studies have been conducted to investigate the effectiveness of these design principles. Although a knowledge base concerning the design of authentic performance assessments is gradually evolving, it remains a complex task to translate the design principles into concrete assessment procedures. The aim of this dissertation was to contribute to the knowledge base concerning issues and measures pertaining to the reliability, generalizability, and validity of performance assessment procedures in order to enable improvement of the methodological quality of such procedures. To realize this, a performance assessment procedure was developed based on design principles from the literature. The performance assessment procedure developed is referred to as 'video portfolios'. These portfolios consist of a mix of sources of evidence that were expected to provide assessors with a complete picture of teachers' competence. In this study, the video portfolios were aimed at measuring the coaching competence of teachers working in senior secondary vocational education. The assessment procedure was designed for this specific teacher competence, because the coaching of students has become an important competence owing to the recent implementation of self-regulated and competence-based education in vocational education. Based on the work of Frederiksen, Sipusic, Sherin, and Wolfe (1998), the main sources of evidence were video episodes representing coaching performance. For this, teachers were filmed on the job during coaching sessions with groups of students. The video episodes represent performance in an authentic context. In order to enable valid scoring and judging of teachers' coaching performance in the video episodes, other sources of evidence were also included in the video portfolios. These sources concerned information about the learning tasks the students worked on during a video episode, information about students' progress with regard to completing the task, the students' backgrounds, the teachers' backgrounds, interviews with the teachers about the decisions underlying their actions, and interviews with students about the perceived impact of the teachers' actions on their work. In the interviews with the teachers, questions were posed with regard to the reasons for coaching, the aims the teachers wished to achieve with the students, the approaches the teachers used, and the extent to which the teachers were satisfied with the results of their coaching. The interviews with the students concerned whether the students felt that the teacher had helped them with a specific topic or problem and whether the support came at the right time. In addition to these sources of evidence, information was added to the video portfolios about educational materials used and students' products that were discussed during the video episodes. Following the development of the performance assessment, trained assessors scored the video portfolios according to a detailed scoring procedure.

The central research question of this dissertation was the following: to what extent are judgments based on video portfolios reliable, generalizable, and valid? In order to answer this question, more specific research questions were formulated and addressed in three studies.

In Study 1, the authentic performance assessment was developed and tested on a small scale. In order to get an indication of the methodological quality of the assessment procedure constructed, two aspects of this were investigated. First, the interrater agreement between assessors was examined as part of the reliability of scores. Second, the utility of the video portfolio assessment procedure with respect to making valid interpretations and judgments was examined. The following research questions were addressed in Study 1:

1a) To what extent do assessors arrive at corresponding scores for video portfolios when judging them using the designed scoring procedure?

1b) Which aspects of the video portfolio assessment procedure stimulate or hinder assessors in making valid interpretations and judgments?

In Study 2, the reliability of assessors' scores using the designed performance assessment was investigated in-depth, and based on a larger sample of assessors. Another important aspect of the methodological quality was also examined in this study: the generalizability of performance scores. The following research questions were addressed:

2a) To what extent are assessors capable of scoring teachers' coaching competence in a reliable way based on a video portfolio?

2b) To what extent can scores assigned to teachers' coaching performance in separate video episodes be generalized to the intended universe of video episodes?

In Study 3, first, the interrater agreement with regard to evidence and arguments underlying numeral scores was investigated. Second, assessors' use of the scoring guide and related conceptual framework was investigated as part of the validity of their scoring processes. The following research questions were addressed:

3a) To what extent do assessors justify their scores assigned to teachers' coaching performance as shown in the video episodes using similar evidence and arguments?

3b) What kind of evidence and arguments do assessors report on score forms?

3c) To what extent do assessors report evidence and arguments that correspond with the scoring guide and related conceptual framework for assessing competent coaching?

1.3 Context of the study: Assessing teachers' coaching competence

1.3.1 Competence-based and self-regulated learning in senior secondary vocational education

As part of this dissertation, an authentic performance assessment procedure was developed based on design principles from the literature. 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 and self-regulated learning in vocational education, coaching has become an important competence domain for teachers. 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). As part of this innovation, a specific project was started called the 'MTS+ project'. This project was implemented in the first and second years of the highest level of vocational education (beroepsopleidende leerweg, niveau 4), in relation to technical studies. This section of vocational education concerns building and construction techniques, electro technology, and mechanical engineering. Students are educated in construction technology, theory of strength of materials, architectural drawing, etc. Students between 16 and 20 years of age enter this type of education. In the MTS+ project, a specific learning environment was developed to foster self-regulated learning by organizing the education around complex tasks that were closely related to tasks people undertake in practice. The complex tasks entailed, for example, designing homes or constructing a dam. Relevant domain-specific knowledge and (to some extent) skills had to be used while students worked on these tasks. The tasks were relatively large projects; students worked on a single task for approximately four weeks. Furthermore, four to six students worked together on one complex task. Since the students did not posses all knowledge and skills needed to fulfill the complex task beforehand, the teachers were expected to coach the students. The teachers were expected, first, to coach the students in performing learning activities that they were not (yet) able carry out on their own, and, second, to coach the students in developing realistic perceptions of professional thinking and acting in practice.

1.3.2 Competent coaching

The development of the assessment procedure started with the construction of an interpretive framework that reflected all relevant aspects of teachers' coaching competence. This conceptual framework was then elaborated on in a concrete scoring guide that included scoring rules, criteria, and performance levels that could be used for scoring and judging video portfolios.

The first step in defining competent coaching was to formulate teacher interventions that could be marked as 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, 2003; 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). 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 four different learning activities: cognitive, meta-cognitive, and affective learning activities, and learning activities related to collaborative learning (Perry, 1998; Perry, Phillips, & Dowler, 2004; Shuell, 1993; Vermunt & Verloop, 1999, Winne & Hadwin, 1998). 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 or 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 related to collaborative learning concern communication, coordination, and the realization of a positive group climate (Johnson & Johnson, 1994; Slavin, 1990).

Once the definition of coaching was determined, competent coaching was defined. Concrete criteria and performance levels for competent coaching were elaborated. In defining competent coaching, a general definition of teachers' competence developed by Roelofs and Sanders (2007) was used as a starting point. According to this definition, teachers' competence is 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. 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, coaching was considered in this dissertation as competent coaching when teachers used coaching interventions that provided students with opportunities to improve their learning activities and their perceptions of practice. Opportunities for improving learning activities can be created through constructive coaching, and opportunities for improving perceptions of practice through practice-oriented coaching. In constructive coaching, the teacher provides just enough support to enable the students to make the step to a higher level in employing learning activities, which they couldn't have made on their own (Vygotsky, 1978). As the improvement in performing a learning activity increases, the support of the teacher decreases until the student can perform the learning activity by him/herself; 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 improvement 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). In practice-oriented coaching, a coach 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.

Four levels of performance were formulated based on the criterion for constructive coaching, and four levels based on the criterion for practice-oriented coaching. For each level, illustrative level descriptors were made. The performance levels indicated the extent to which teachers' behavior led to positive consequences for students. The descriptors were expected to assist assessors in making relevant considerations and in deciding which performance level was matched by the coaching performance observed.

1.3.3 Scoring procedure

The assessors were expected to score the video portfolios according to a detailed scoring procedure. In this procedure, for each video episode, the assessors started by collecting specific evidence pertaining to teachers' questions and feedback that did or did not provide the opportunity for students to improve their performance of learning activities and perceptions of practice. The assessors then used the specific evidence to build a judgment of the performance in the whole video episode, based on four distinguished performance levels. Subsequently, the assessors formed an overall judgment about teachers' coaching competence based on the performance across the video episodes. In this judgment, the four performance levels were also used. The assessors were urged to follow the steps of the scoring procedure in detail.

1.3.4 Assessor training

Assessor training has emerged in the literature as a prerequisite for accurate ratings in performance assessment (Day & Sulsky, 1995; Stamoulis & Hauenstein, 1993; Uhlenbeck, 2002; Woerh & Huttcuff, 1994). For that reason, an assessor training course was set up to prepare the assessors for scoring and judging the final video portfolios. Four training sessions were developed in which the assessors learned to use and apply the constructs from the conceptual framework and the detailed scoring procedure in a systematic and consistent way. 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 the assessors received feedback. During the training, assessors were corrected when they deviated from the scoring procedure. Another important goal of the training was to make assessors aware of rating errors; they were urged to correct those errors immediately if they occurred. Special attention was given to errors concerning inappropriate emphasis on specific evidence or arguments, selective observation, inconsistencies in assessors' scoring, halo-effect, horn-effect, and central tendency (Aronson, Wilson, & Akert, 2007).

1.4 Relevance of the study

The theoretical relevance of this dissertation lies in its contribution to the knowledge base concerning issues and measures in performance assessments pertaining to reliability, generalizability, and validity. The studies presented in this dissertation were aimed at investigating the design of performance assessments in relation to methodological issues that specifically pertain to assessors' scoring (scoring inference) and to the generalization of an observed score on an assessment task to a universe score (generalization inference). The studies were aimed at providing greater insight into the extent to which the design principles proposed in the literature actually lead to reliable and valid performance assessments. Another goal of the studies was to provide greater insight into the magnitude of threats to reliable and valid scoring and into the generalizability of scores across assessment tasks. In brief, this research was aimed at expanding insights into the occurrence of threats such as the impact of assessors' personal beliefs, assessors' biases, selective observation, and task specificity, as well as into measures in the design of performance assessments that can reduce these threats.

The practical relevance of this dissertation lies in its contribution to insights into the methodological opportunities and restrictions of performance assessment. These insights can be used in the design of performance assessments and in the use of these instruments in teacher education and schools. It is expected that authentic performance assessment will be used more frequently in the future to make different (high-stakes) types of decisions about, for example, the selection of educational personnel, admission to teacher education or continuation of the course in teacher education, differential payment, and licensure of educational personnel. In the event of such situations it is important to be knowledgeable about the methodological quality of performance assessments.

The practical relevance of this research lies also in the development of an assessment procedure for assessing teachers' coaching competence in senior secondary vocational education. The performance assessment developed in this study can be used to determine to what extent teachers use the relatively new, but relevant coaching competences they need for teaching in the new learning environment aimed at competence-based and self-regulated learning.

1.5 Outline of the study

In this dissertation, three studies are presented in which the reliability, generalizability, and validity of performance assessments were examined in different manners. Table 1 provides an overview of the design of the three studies. In the columns, the following information is presented: research questions; the inference (in Kane's chain of inferences) that was evaluated; the type of data collected; the sample of assessors, video episodes, or score forms that were included in the analyses; and the analyses done in order to answer the research question.

In Chapter 2, the first study is presented. In this study, the assessment procedure was designed and tested on a small scale. Two important aspects of the reliability and validity of the video portfolios were investigated: the interrater agreement between assessors, and aspects in the design of the video portfolios that stimulated or hindered assessors in making valid interpretations and judgments. This study was focused on the first inference of Kane's chain of inferences: the scoring inference. To investigate this inference, scores assigned to video episodes and overall scores were collected, and the interrater agreement was determined. Furthermore, a semi-structured interview was carried out with all assessors in order to obtain information on aspects of the assessment procedure that stimulated or hindered assessors in making valid interpretations and judgments.

In Chapter 3, the second study is presented. In this study, the reliability of assessors' scoring was investigated in-depth and based on a larger sample of assessors. This part of the study was focused on the scoring inference. Scores assigned to video episodes and overall scores were collected, and several qualitative analyses were conducted. These analyses concerned the examination of tendencies in assessors' assigned scores, interrater agreement, and generalizability across assessors. In the second study, another aspect of the methodological quality of the video portfolios was also examined: the generalizability of scores across video episodes. This part of the study was focused on the evaluation of the generalization inference. Several analyses were conducted in order to examine this aspect. A ranking order was made from video episodes that elicited the most similar scores to video episodes that elicited the most varying scores. It was expected that especially the video episodes that provoked the most varying scores would be a threat to the generalizability across video episodes.

Furthermore, it was determined for each video episode to what extent the score assigned to it matched the scores assigned to the other video episodes.

In Chapter 4, the last study is presented. In this study, the validity of assessors' scoring was investigated. The emphasis was once more on the scoring inference. However, in contrast to the previous studies, the analyses were focussed on the evidence and arguments assessors used to justify the scores assigned. Most validity and reliability research focuses on the technical soundness of the assessment procedures. However, statistics lack information about the process of scoring and the actual use of the scoring rules by raters (Linn, 1994; Messick, 1995; Moss, 1994; Van der Schaaf, Stokking, & Verloop, 2005). For that reason, qualitative analyses were used in this study to enable more thorough investigation of the validity of assessors' scoring processes.

Table 1.1 Design of the three studies

	Research questions	**Inference evaluated**	**Method**		**Analysis**
			Data collected	**Sample**	
Chapter 2	To what extent do assessors arrive at corresponding scores for video portfolios when judging them using the designed scoring procedure?	Scoring inference	Assigned scores reported on score forms	6 assessors, 28 videos, and 8 overall scores	Quantitative analysis concerning rater agreement
	Which aspects of the video portfolio assessment procedure stimulate or hinder assessors in making valid interpretations and judgments?	Scoring inference	Semi-structured interview concerning assessors' experiences in applying the assessment procedure	6 assessors	Qualitative content analysis on interview data concerning: content video portfolios, judging and interpreting videos, criteria and performance levels, and scoring procedure

Table 1.1 Design of the three studies (Continued)

	Research questions	Inference evaluated	Method		Analysis
			Data collected	Sample	
Chapter 3	To what extent are assessors capable of scoring teachers' coaching competence in a reliable way based on a video portfolio?	Scoring inference	Assigned scores reported on score forms	12 assessors, 38 videos, and 11 overall scores	Quantitative analysis concerning rater agreement and scoring tendencies
	To what extent can scores assigned to teachers' coaching performance in separate video episodes be generalized to the intended universe of video episodes?	Generalization inference	Assigned scores reported on score forms	12 assessors and 38 videos	Quantitative analysis concerning correlations between scores and rest scores
Chapter 4	To what extent do assessors justify their scores assigned to teachers' coaching performance as shown in the video episodes using similar evidence and arguments? What kind of evidence and arguments do assessors report on score forms? To what extent do assessors report evidence and arguments that correspond with the scoring guide and related conceptual framework for assessing competent coaching?	Scoring inference	Evidence and arguments reported on score forms that were used to justify a score	126 score forms from 12 assessors	Qualitative and quantitative content analysis of evidence and arguments reported on score forms

Finally, Chapter 5 presents the general conclusions and discussion based on the findings of the three studies described in the previous chapters, and offers suggestions for future research and practical implications for the design of valid and reliable authentic performance assessment procedures.

References

Aronson, E., Wilson, T.D., & Akert, R.M. (2007). *Social psychology* (5th ed.). Amsterdam: Pearson Education Benelux BV.

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.

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.

Borsboom, D., & Mellenbergh, G.J. (2004). The Concept of Validity. *Psychological Review, 111*(4), 1061-1071.

Brennan, R.L. (1998). Misconceptions at the intersection of measurement theory and practice. *Educational Measurement: Issue and Practice, 17*(1), 5-9.

Brennan, R.L. (2000). Performance assessments from the perspective of generalizability theory. *Applied Psychological Measurement, 24*(4), 339-353.

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* (pp. 453-494). Hillsdale, NJ: Erlbaum.

Crocker, L. (2003). Teaching for the test: Validity, fairness, and moral action. *Educational Measurement: Issue and Practice, 22*(3), 5-11.

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.

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

Dunbar, S.B., Koretz, D.M., & Hoover, H.D. (1991). Quality control in the development and use of performance assessments. *Applied Measurement in Education*, *4*, 289-303.

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

Eraut, M. (1994). *Developing professional knowledge and competence.* London: Falmer Press.

Fredriksen, R.F., & Collins, A. (1989). A systems approach to educational testing. *Educational Researcher, 19*(9), 27-32.

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.

Gonzi, A. (1994). Competency based assessment in the professions in Australia. *Assessment in Education 1*(1), 27-45.

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

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

Kane, M.T. (2004). Certification testing as an illustration of argument-based validation. *Measurement, 2*(3), 135-170.

Kane, M.T. (2006). Validation. In R.L. Brennan (Ed.), *Education Measurement* (4th ed.). Westport: Praeger Publishers.

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

Linn, R.L. (1994). Performance assessment. Policy promises and technical measurement standards. *Educational Researcher, 23*(9), 4-14.

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

Linn, R.L., & Burton, E. (1994). Performance-based assessment: Implications of task-specificity. *Educational Measurement: Issues and Practice, 13*(1), 5-15.

Lissitz, R.W., & Samuelson, K. (2007). A suggested change in terminology and emphasis regarding validity and education. *Educational Researcher*, *36*(8), 437-448.

Messick, S. (1989). Validity. In R.L. Linn (Ed.), *Educational measurement* (3^{rd} ed.). New York: MacMillan.

Messick, S. (1994). The interplay of evidence and consequences in the validation of performance assessments. *Educational Researcher, 23*(2), 13-23.

Messick, S. (1995). Validity of psychological assessment. Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist, 50*(9), 741-749.

Miller, M.D., & Linn, R.L. (2000). Validation of performance assessments. *Applied Psychological Measurement 24*(4), 367-378.

Mislevy, R.J. (2007). Validity by design. Comments on Lissitz and Samuelson. *Educational Researcher*, *36*(8), 437-448.

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.

Onstenk, J. (2000). *Op zoek naar een krachtige beroepsgerichte leeromgeving: fundamenten voor een onderwijsconcept voor de bve-sector (A search for powerfull 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.

Ruiz-Primo, M.A., Baxter, G.P., & Shavelson, R.J. (1993). On the stability of performance assessments. *Journal of Educational Measurement, 30*, 41-53.

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?* (pp. 15-45). Newbury Park, CA: Corwin Press.

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

Shavelson, R.J., Baxter, G.P., & Gao, X. (1993). Sampling variability of performance assessments. *Journal of Educational Measurement, 30*, 215-232.

Shephard, L.A. (1993). Evaluating test validity. *Review of Research in Education, 19*, 405-450.

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

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.

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

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.

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.

Chapter 3

Reliability and generalizability of performance judgments based on a video portfolio[2]

Abstract

Authentic teacher assessments are increasingly developed and used in practice. An important issue in designing authentic performance assessments is how the reliability and validity of these assessments can be guaranteed. In the literature, several design principles are discussed that should contribute to more reliable and valid assessments, such as increasing the number of assessors and assessment tasks in the assessment, standardizing assessment tasks, and using high-fidelity tasks in the assessments. However, not much empirical evidence is available that proves that these principles really contribute to reliable and valid assessments. The aim of this research was to find out whether these design principles lead to reliable and valid assessments. Previous to this study, an authentic performance assessment was constructed based on the design principles (see chapter 2). The assessment constructed can be used for assessing teachers' coaching competence in the context of senior secondary vocational education. Video recordings of teachers' coaching performance in the classroom are the main elements of the assessment procedure constructed. Additional data sources were included that provided information about the contexts of the videotaped coaching situations. This combination of video recordings and context information is called a 'video portfolio'. After the construction of the video portfolios, their validity was determined by answering the following research questions: (a) To what extent did the assessors score teachers' coaching competence in a reliable way based on the video portfolios? (b) Can scores assigned to separate video episodes be generalized to the intended universe of video episodes? In order to answer these research questions, twelve assessors were asked to score four video portfolios. Scorecards were gathered and several analyses were performed on the scores assigned in order to get an indication of the interrater agreement and of the generalizability of scores across video

[2] This chapter has been published in adapted form as:
Bakker, M., Sanders, P., Beijaard, D., Roelofs, E., Tigelaar, D., & Verloop, N. (2008). De betrouwbaarheid en generaliseerbaarheid van competentiebeoordelingen op basis van een videodossier. *Pedagogische Studiën 85*(4), 240-260.

This chapter has also been submitted in adapted form as:
Bakker, M., Sanders P., Beijaard, D., Roelofs, E., Tigelaar, D., & Verloop, N. Reliability and generalizability of performance judgments based on a video portfolio.

episodes. It appeared that the design principles went together with positive results concerning assessors' scoring. An acceptable to high level of interrater agreement was found for scores assigned to video episodes, and a high level of interrater agreement was found for the overall scores assigned. Furthermore, there are strong indications that the design principles went together with positive results concerning the generalizability of scores assigned across video episodes. Except for one assessment scale (coaching with regard to affective learning activities), an acceptable to high level of similarity was found between scores assigned to a video episode and the average of the scores assigned to the other video episodes on the assessment scale.

3.1 Introduction

Much attention is currently given to the design and use of authentic performance assessments. These assessments are used to gain insight into the level of teacher competence (summative assessment) as well as to provide a starting point for further professional development (formative assessment). A knowledge base has gradually emerged pertaining to the assessment of teacher competence. Contemporary researchers ascertain that to ensure that the assessment can be used for summative as well as formative assessments, a mix of evidence sources should be used, collected in authentic task situations (Darling-Hammond & Snyder, 2000; Dwyer, 1998; Gipps, 1994; Haertel, 1991).

Typically, in performance assessments, the teacher is asked to perform, produce, or create something over a sufficient duration of time to permit evaluation of either the process or the product of performance, or both. Examples can be found in Haertel (1991), Peterson (2001), and Uhlenbeck (2002), and entail, for instance, use of teacher work samples, teacher portfolios, peer review of materials, systematic observation, reflective interviews, performance exercises as lesson planning, and review of students' assignments. In sum, performance assessments consist of multiple tasks to be carried out by respondents (Kane, 2004). In addition, a central role is played by the assessors who interpret the performance of the respondents. When the validity of a performance assessment is to be investigated, respondents, tasks, and assessors have to be taken into account.

Kane (2006) developed a procedure by which a (performance) assessment can be validated. In his validity argument-based approach, he states that the validity of an assessment can be investigated by evaluating the chain of inferences that takes place when the outcomes of a performance assessment are interpreted. Three inferences form the heart of the validity argument: (1) reliable and valid scoring of performance by assessors, (2) generalization from the score observed on an assessment task to a universe score, (3) extrapolation of assessment results to practice. In a thorough validity investigation, the tenability of all three inferences should be examined.

Until recently, researchers focussed on interrater reliability as an indication of a reliable assessment (Dunbar, Koretz, & Hoover, 1991). The scoring of a teacher's performance by assessors was found to be a difficult task (Gipps, 1994; Moss, 1994). An explanation for this is that, in performance assessments, complex and open tasks are used that are often situated in varying contexts. Respondents can react to those assessment tasks in many different ways, and it is not easy for assessors to score the varying information that results in a consistent way. Especially selective observations, personal prejudices, and biases are serious threats to the reliability and validity of the scoring process (Gipps, 1994; Moss, 1994).

Currently, more attention is given to the extent to which the assessment tasks can be generalized to a broader domain of assessment tasks. In addition, more attention is given to the question of whether the sample of assessment tasks can be seen as a representation of the construct to be measured. In other words, it is examined whether the scores on the sample of assessment tasks can be extrapolated to performance in daily practice. A problem in constructing a representative sample of assessment tasks is that complex and open-ended assessment tasks are time consuming. Only a restricted number of tasks can be included in the performance assessment, so it may turn out to be difficult to extrapolate the performance measured to performance in daily practice (Brennan, 2000; Dunbar, Koretz, & Hoover, 1991; Linn, Baker, & Dunbar, 1991; Linn & Burton, 1994; Miller & Linn, 2000; Ruiz-Primo, Baxter, & Shavelson, 1993; Shavelson, Baxter, & Gao, 1993).

Several design principles can be used that can ensure the tenability of the scoring, generalization, and extrapolation inference. Examples of such design principles are increasing the number of assessors, standardising assessment tasks, and using authentic assessment tasks. The aim of this study was to examine the extent to which these design principles actually contribute to valid and reliable performance assessment. Previous to this study, a performance assessment procedure was developed, based on several design principles for valid and reliable scoring, generalization, and extrapolation. The general design principles are discussed in section 3.2. The actual design measures applied to the performance assessment constructed are discussed in section 3.3. The performance assessment was aimed at assessing teachers' coaching competence in the context of senior secondary vocational education. As a result of the implementation of self-regulated learning in Dutch vocational education, teachers are expected to coach their students while they work independently on complex, job-related tasks (Moerkamp, De Bruijn, Van der Kuip, Onstenk, & Voncken, 2000; Onstenk, 2000). The teachers' coaching performance is assessed using the video portfolio method. Based on the work of Fredriksen, Sipusic, Sherin, and Wolfe (1998), the main components of a video portfolio are video episodes of teachers' coaching performance in key situations in the classroom. In order to interpret and judge teachers' performance in a valid way, supporting data sources were included in the video portfolios that outlined the contexts in which the coaching took place. The content of a video portfolio and the scoring procedure are discussed in detail in section 3.3. Four video portfolios were constructed and subsequently scored by twelve trained assessors. Afterwards, the validity of the method was investigated using the chain of inference approach mentioned above.

3.2 Validity and reliability in scoring, generalization, and extrapolation

Reliability is defined as the extent to which the results of an assessment can be repeated. It entails the question of whether assessment results will vary when the assessment is repeated under the same conditions. In recent decades, the definition of validity has undergone some changes. Three perspectives on validity have been distinguished: criterion validity, content validity, and construct validity. Criterion validity refers to the relationship between the test score and an external criterion that is viewed as a direct measurement of the characteristic to be measured. Content validity refers to the extent to which the measurement is representative of the domain to be measured. Construct validity concerns the extent to which the construct (or characteristic) to be measured, is measured. Nowadays, this traditional classification of validity receives less support. Construct validity is now seen as a term that also covers criterion validity and content validity (Messick, 1989). Validity is seen as "an integrated evaluative judgment of the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of inferences and actions based on test scores or other modes of assessment" (Messick, 1989, p.13). Although some objections can be made against this definition of validity, like that it is very broad (Borsboom & Mellenbergh, 2004), it has been generally accepted since the eighties.

The validity of an assessment procedure can be investigated systematically by examining the chain of three inferences (Kane, 2006). These three inferences are scoring, generalization, and extrapolation. They are shown in Figure 3.1.

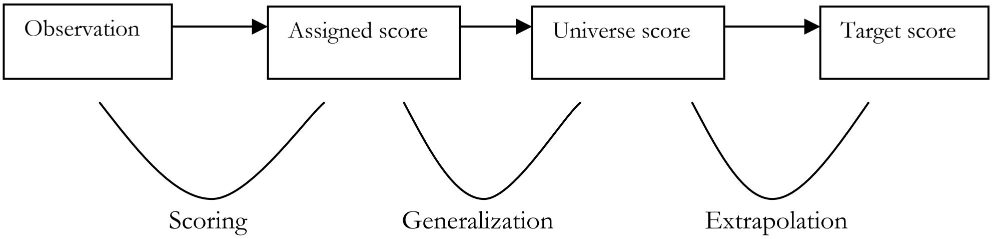


Figure 3.1 Chain of inferences for a validity argument regarding performance assessments

Scoring of performance

The first inference from the chain pertains to the scoring of the performance of respondents by the assessor: are the assessor's interpretations and judgments of the performance valid and reliable? Especially the influence of personal characteristics on judgments is a serious threat to the tenability of the first inference regarding scoring, such as selective observation, biases, and personal prejudices (Gipps, 1994; Moss, 1994). Several factors can influence the tenability of the scoring inference. First, judgments are more valid and reliable when appropriate criteria, performance levels, and scoring rules are used during the scoring process and when assessors are capable of applying these in a consistent way. Assessor training has a positive influence on the application of criteria, standards, and scoring rules (Day & Sulsky, 1995; Stamoulis & Hauenstein, 1993). Second, a large number of assessors contributes generally to more reliable scoring (Kane, 2006). When multiple assessors judge a performance, the personal influence on the judgment of individual assessors decreases, so that the scores assigned are more accurate. Third, it appears that assessors score a performance in a more consistent way when all respondents perform the same assessment tasks instead of different tasks. This mainly leads to more reliable judgments (Crooks, Kane, & Cohen, 1996).

Generalization across assessment tasks

In determining the tenability of the second inference, the following question is relevant: does the score obtained based on the assessment tasks represent the score that a respondent would have achieved if he or she had accomplished all possible tasks used to measure the construct to be measured? When examining this inference, it should be investigated whether a respondent would have received a different assessment result if he or she had accomplished other assessment tasks. This concerns the question of whether the sample of assessment tasks used in the assessment is representative for the universe of assessment tasks. A universe of assessment tasks refers to the collection of assessment tasks out of all possible tasks that are appropriate to measure the construct at hand (Sanders, 1998). Particularly this second inference seems problematic in performance assessment. Respondents show very divergent performances while performing different tasks, even when the tasks are from the same domain. A measure to overcome this problem is to standardize assessment tasks. In standardizing assessment tasks, the aim is to create tasks that call upon the same characteristic every time, so that the agreement in assigned scores between the tasks will be large. When the agreement on different tasks is large, it is better possible to generalize the scores to a universe score. Furthermore, it is easier when using standardized assessment tasks to formulate detailed scoring rules, and it is easier for assessors to score the performance in a consistent way (Brennan, 2000; Kane, 2006).

Extrapolation to performance outside the assessment context

In examining the third inference, it is investigated to what extent it is possible to extrapolate the performance as measured in the assessment to performance outside the assessment context. A design principle used to enable extrapolation to performance outside the assessment context is the use of so-called 'high-fidelity tasks' (Kane, 2006). These tasks measure the characteristic in a very direct way. However, high-fidelity tasks are often complex and open-ended tasks that are hard for assessors to score. Furthermore, these tasks are very time consuming, so that for reasons of practical feasibility, only a restricted number of tasks can be included in an assessment. As a result of the restricted number of assessment tasks, it can be hard to establish a representative sample to enable extrapolation to performance outside the assessment context. Especially the use of a large number of assessment tasks has a considerable positive effect on extrapolation to performance outside the assessment context (Dunbar, Koretz, & Hoover, 1991; Ruiz-Primo, Baxter, & Shavelson, 1993). This remains a difficult issue in performance assessment; no clear-cut solution is at hand.

In this study, two of the three inferences of the model introduced by Kane (2006) were investigated. The following research questions were answered:

- To what extent are assessors capable of scoring teachers' coaching competence in a reliable way based on a video portfolio?
- To what extent can scores assigned to the coaching performance in separate video episodes be generalized to the intended universe of video episodes?

In answering the first research question, the investigation was restricted to an examination of the reliability of the performance scores assigned. In a subsequent study (see chapter 4), the scoring process, including the construct relevance of assessors´ considerations and arguments regarding teachers´ performances, were examined in more detail. For answering the second research question, usually an generalizability study is conducted. However, because the construction of the video portfolios according to design principles was a complex and time consuming process, it was not possible to establish a substantial sample of video portfolios that is needed to determine the generalizability of scores based on a generalizability study. Therefore, other methods are used to obtain an indication of the generalizability of scores. The third inference (extrapolation to performance outside the assessment context) was not investigated in this study. To investigate this inference, a job analysis would be needed to show what coaching situations occur in practice, and how often. So far, no job analysis is available. For that reason, we decided not to include investigation of this inference in this study.

3.3 Method

3.3.1 Design of the performance assessment procedure

Based on a literature study in the field of supporting self-regulated learning (Boekaerts, 1999; Boekaerts & Simons, 1995; Bolhuis, 2000; Butler & Winne, 1995) and on observations in practice, coaching was defined as supporting learning activities that students can not (yet) carry out on their own. Typical interventions that can be used by teachers to support or coach students in carrying out learning activities are asking questions and providing feedback (Boekaerts & Simons; 1995; Butler & Winne, 1995). These coaching interventions can be used to support four different types of learning activities. Firstly, students' learning activities that concern activities to process subject matter and that lead to learning outcomes in terms of changes in students' knowledge base and skills (cognitive learning activities). Secondly, learning activities that 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 (affective learning activities). Thirdly, learning activities that concern thinking activities which 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 (meta-cognitive learning activities). Finally, learning activities that pertain to collaboration with other students. Knowledge about coaching for self-regulated learning, encompassing the first three learning activities mentioned, is based on instructional theories elaborated by Shuell (1993), Vermunt and Verloop (1999), and Winne and Hadwin (1998). Coaching in the fourth learning activity is based on theories about collaborative learning (Johnson, & Johnson, 1994; Slavin, 1990).

Following this, an assessment scale was constructed to enable expression of the level of performance. The starting point in constructing the performance levels was the definition of competent teaching by Roelofs and Sanders (2007). They see competent teaching as being able to make appropriate and deliberate decisions in a specific context, based on a personal knowledge base, which results in behavior that contributes to desired consequences. Competent coaching was then defined. In this study, competent coaching was defined as constructive coaching. Constructive coaching entails that the teacher uses coaching interventions that provide students with opportunities and stimulate them to improve the self-regulating learning activities described above. In constructive coaching, the teacher provides just enough support so that the students can make the step to a higher level in employing learning activities, which they couldn't have made on their own (Vygotsky, 1978). As the performance of a learning activity improves, the support of the teacher decreases until the student can perform the learning activity by him/herself; this is referred to in the literature as 'fading' (Collins, Brown, & Newman, 1989). Table 2.2 in chapter 2 presents the performance levels of (non-) constructive.

Video portfolios

The performance levels were used to score and judge the video portfolios. A video portfolio consists of a mix of information sources that are expected to provide assessors with a complete picture of teachers' coaching competence. The main sources of evidence consist of video episodes that represent teachers' coaching performance in key situations. In order to enable the assessors to score and judge the teachers' coaching performance in the video episodes in a valid way, information about the context was added: information about the learning task the students worked on during a video episode; information about students' progress in completing the task; information about students' backgrounds; information about the teachers' backgrounds; interviews with the teachers about the decisions underlying their actions; and interviews with student(s) about the perceived impact of teachers' actions on their work. The interview with the teachers concerned questions about the reasons for coaching, the aims the teacher wished to achieve with the students, the approach the teacher used, and the extent to which the teacher was satisfied with the results of his or her coaching. The interview with the students was aimed at examining whether a teacher support with regard to a specific topic or problem helped them, and whether the support came at the right time.

Scoring procedure

Twelve assessors scored the video portfolios according to a detailed scoring procedure. The scoring procedure is presented in Table 2.5 in chapter 2 and the score forms used during the scoring are presented in Appendix 2 and 3. In the scoring procedure presented in chapter 2 and on the score forms in Appendix 2 and 3, also instructions are included for scoring practice-oriented coaching. In this study, assessors were asked to score teachers' coaching performance only for constructive coaching. This was decided based on the findings in study 1, which showed that practice-oriented coaching could not be scored in a valid way based on the video portfolios constructed.

3.3.2 Measures to achieve reliable and valid scoring

Scoring guide and related conceptual framework

In the design of the assessment procedure, several measures were taken to achieve scoring that was as reliable and valid as possible. In order to reduce the impact of personal biases and beliefs on scores, and to minimize the occurrence of selective observation and judging according to personal constructs (DeNisi, Cafferty, & Meglino, 1984; Feldman, 1981; Landy, & Farr, 1980; Van der Schaaf, Stokking, & Verloop, 2005), a scoring guide and a related conceptual framework containing relevant concepts and criteria were constructed. In this study, the assessors were provided with a scoring guide and a related conceptual framework pertaining to competent coaching. Moreover, the assessors were trained in using this scoring guide.

Theory and practice

The construction of the scoring guide and conceptual framework was started with a literature study. The literature-based framework was presented to and discussed with teachers working in senior secondary vocational education. Observations were made in order to obtain information about the kinds of coaching interventions teachers use in practice. Based on these interviews and observations, the literature-based scoring guide was refined and adjusted to the context of senior secondary vocational education. As a result of adjusting the framework to the context in vocational education, it was expected that the scoring guide would lead to more appropriate criteria for competent coaching. This should lead to a valid scoring guide, which should contribute to more valid scoring by assessors.

Concrete examples of coaching interventions

During the construction of the scoring guide, examples of coaching interventions were collected that teachers used in practice. It was expected that these examples would help assessors in identifying relevant coaching interventions (Frederiksen, Sipusic, Sherin, & Wolfe, 1998). As a result of being given concrete examples, assessors were expected to know better what to look for in a video episode showing a coaching performance. The inclusion of concrete examples in the scoring guide was expected to contribute to higher interrater agreement.

Use of performance levels

In order to enable assessors to score the coaching performance, the scoring guide included four performance levels. For each level, illustrative level descriptors were constructed. The descriptors contained information about teachers' behavior and consequences for students that were specific to that level of performance. The level descriptors were expected to assist assessors in making relevant considerations and decisions. Furthermore, the level descriptors were expected to assist assessors in scoring performance in different contexts in a consistent way, so that higher interrater agreement could be reached.

Scoring procedure

The scoring guide contained a detailed scoring procedure. In this scoring procedure, assessors started by scoring specific aspects of the performance according to guidelines and criteria. Assessors then used these scores to assign an overall score for the whole performance. Because the scoring procedure was structured using (detailed) guidelines, it was expected that assessors would have little room to base their judgments on their personal biases and beliefs, which should result in more objective and reliable judgments (Klein, & Stecher, 1998). The scoring procedure was elaborated along with measures that were expected to lead to more valid interpretation processes, as described by Moss, Schutz, and Collins (1998) and Schutz and Moss (2004). The first measure was that assessors were urged to consider all available evidence and to check afterwards whether they had based the score assigned on all available evidence. The second measure was that assessors should actively seek counter-evidence in order to reduce the impact of construct under-representation. In the scoring procedure, assessors were urged to search for coaching interventions demonstrated by the teacher that did provide opportunities for students as well as interventions that did not. The third measure was that assessors should challenge one another's interpretations, so that the acceptability and tenability of the interpretations would be critically checked. In that way, the impact of selective observation, personal points of view, beliefs, and opinions should be reduced as much as possible. In order to provide a chance to exchange interpretations and judgments with another assessor, a discussion phase was included in the scoring procedure (step 4).

Assessor training

Assessor training has emerged as a prerequisite for accurate ratings in performance assessment (Day & Sulsky, 1995; Stamoulis & Hauenstein, 1993; Uhlenbeck, 2002; Woerh & Huttcuff, 1994). For that reason, an assessor training course was set up to prepare assessors for scoring and judging video portfolios. A series of four training sessions, each lasting half a day, was developed. The sessions were aimed at training assessors to use the conceptual framework and the scoring method in a systematic and consistent way.

During the assessor training, video episodes that were not included in the video portfolios were observed and discussed. The scoring method was introduced and applied step by step in practice. The following assessor skills were addressed:

- 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 coaching;
- 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 their training, the assessors were corrected when they deviated from the scoring procedure. Another aim of the training was to make assessors aware of rating errors. Any scoring error that occurred was corrected immediately. Special attention was given to errors concerning an inappropriate emphasis on specific evidence or arguments, selective observation, inconsistencies in assessors' scoring, halo-effect, horn-effect, and central tendency (Aronson, Wilson, & Akert, 2007).

Organization and arrangement of evidence

In order to ensure validity and reliability in assessors' interpretations and judgments, three measures were taken. First, a professional video production company recorded the videos. Three cameras and three microphones were used to record all teacher and student activities at the same time. The starting point was that all interactions between teacher and student(s) would be clearly perceptible for assessors, to ensure that no evidence would be lost. Second, in addition to video episodes, supporting sources of information that outlined the coaching context were included in the video portfolios. It appears that assessors need this information to be able to decide on the level of the coaching performance shown in the video episodes (Heller, Sheingold, & Myfords, 1998; Schutz, & Moss, 2004). Third, the video episodes and context information were visually ordered in a multi-media environment, to enable assessors to evaluate all available evidence in coherence.

3.3.3 Measure to generalize across video episodes

To enable generalization of scores assigned to teachers' coaching performance in a particular video episode to the universe of video episodes, specific video episodes were selected. Although the video episodes represent very authentic teacher performance, it was attempted to standardize the videos by selecting only video episodes that concern a key situation. A key situation is a coaching situation in which students need support in carrying out a specific learning activity to complete the complex task they are working on. It is a situation that is expected to provide valuable evidence of teachers' coaching competence.

3.3.4 Measures to extrapolate to performance outside the assessment context

As mentioned earlier in this study, the measures applied in the assessment procedure in order to extrapolate to performance outside the assessment context were not evaluated in this study. Nevertheless, the measures applied are described in this section. In the video portfolio performance assessment, high-fidelity tasks were used to measure teachers' coaching competence in a very direct way. The high-fidelity tasks were actual coaching tasks that teachers carried out in their classrooms, as a result of emerging learning needs on the part of the students. From all recordings made in the classroom, key situations were selected for inclusion in the video portfolio. In order to be able to extrapolate to teachers' coaching competence outside the assessment context, it was important to create a sample of coaching situations that represented coaching situations that would occur in practice. To establish variation in the video episodes, the video episodes of different key situations were selected on the basis of the following criteria: the sample should contain key situations spread across the four weeks that students worked on one complex task, and covering all stages of learning that might take place. In addition, the sample should contain video episodes that concerned coaching in all the different learning activities. Another important factor in creating a sample of video episodes is the number of video episodes to be included in the video portfolio. The larger the number of video episodes included in the portfolio, the better can be extrapolated to coaching competence outside the assessment context. However, practical feasibility also plays a role here. Assessors can only score a restricted number of video episodes within a reasonable amount of time. Thus, an important consideration is how many video episodes should be included in order to be able to extrapolate, which can also be scored within a reasonable amount of time. In this study, we included ten video episodes in a video portfolio.

3.3.5 Participants

With the technical assistance of a video database specialist, the researchers constructed video portfolios of four teachers working in senior secondary vocational education. The four teachers (one female and three males) worked as coaches in the building technology section and had one to two years' experience in coaching students. They had two different responsibilities. Two of the four coaches coached students mainly in cognitive, meta-cognitive, and affective learning activities (job profile 1); the other two coached the students mainly in meta-cognitive and affective learning activities, and learning activities related to collaborative learning (job profile 2). In the video portfolios constructed, the teachers' responsibilities were taken into account; video episodes were selected that matched their specific job responsibilities as described above.

The video portfolios were scored and judged by twelve trained assessors, who were from the same discipline and had an equal amount of experience in coaching students. Six of the twelve assessors worked at the same school as the teachers recorded in the video portfolios. The other six assessors were from another school.

3.3.6 Data collection

After the four training sessions, the assessors scored the four video portfolios independently. They assigned a score for constructive coaching to the coaching performance in each video episode, corresponding to one of the four levels of coaching competence. They then assigned overall scores, also using the scale with the four performance levels. For coaches with job profile 1, three overall scores were assigned: an overall score for coaching in (a) cognitive, (b) meta-cognitive, and (c) affective learning activities. For coaches with job profile 2, also three overall scores were assigned: an overall score for coaching in (a) meta-cognitive, (b) affective, and (c) learning activities concerning collaboration. The assessors were asked to weigh the scores assigned to the separate video episodes in order to arrive at an overall score. After assigning scores independently, assessors discussed their individually assigned (overall) scores in pairs. Assessors were free to adjust their original scores based on the discussion. Score forms containing the scores assigned were collected.

3.3.7 Analysis: Assessors' scoring

In order to investigate the reliability of the assessors' scoring, several analyses were conducted. First, tendencies in the scores assigned by the assessors were examined. These analyses were carried out in order to determine whether the assessors scored the different teachers equally leniently or severely, and to get an overview of the assessors who assigned extreme lenience and extreme severity. The average scores assigned to the coaching performances across the video episodes in the video portfolios were determined for each assessor and each teacher. The average scores assigned by each assessor to each teacher were visualized in a chart. When the lines in the chart are parallel to each other, the assessors were equally lenient or severe for all teachers. When the lines in the chart are not parallel to each other, the assessors were more lenient or more severe in judging some of the teachers. This analysis was also conducted for the overall scores assigned.

In a second analysis the interrater agreement on assigned scores was examined. In this type of analysis it is common to exclude the assessors who assigned the most extreme scores. For that reason the analyses were conducted twice: once including the extreme assessors and once excluding them. In this study, the frequency of cases where 50% or more of the assessors assigned the exact same (overall) score was used as an indication of agreement. The Gower coefficient was also used as an indication of interrater agreement with regard to assigned (overall) scores. A generalizability coefficient is usually used as an indicator for rater agreement. Variance components of respondents, assessors, assessment tasks, and interaction effects between these facets are estimated in a generalizability study. 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 based on absolute differences between assigned scores. In addition, the range of the assessment scale is taken into account. The coefficient is not only based on the cases where assessors assign the exact same score to a performance, but also takes into account the absolute distance between the assigned scores on the assessment scale when assessors do not assign the same score.

The formula for determining a Gower coefficient is the following:

$$G_{xy} = 1- \{ \Sigma \mid X_i – Y_i \mid / nR \}$$

Xi and Yi in the formula represent the scores assigned by two assessors. The number of objects judged is represented by n, and the range of the assessment scale by R (Zegers, 1989). The Gower coefficient ranges from 0 (no agreement between assessors) to 1 (perfect agreement between assessors). A Gower coefficient from 0 to 0.65 is perceived as low, a Gower coefficient between 0.65 and 0.85 is perceived as acceptable, and a Gower coefficient between 0.85 and 1 is perceived as high. As the formula indicates, the Gower coefficient is used to compare the scores assigned by two assessors. In the analyses conducted in this study, a Gower coefficient was determined for every possible pair of assessors. The Gower coefficients reported in section 4 are average Gower coefficients across all assessor pairs.

The findings of the third analysis enabled us to get an indication of the minimum number of assessors that should be involved in a performance assessment in order to attain reliable scores. This is an important issue. In this study, twelve assessors were involved in scoring the video portfolios; in practice, however, it is often impossible to involve such a large number of assessors, for reasons of time and costs. If generalizability of scores across assessors increases, then fewer assessors are needed to reach an acceptable level of agreement. In this analysis, it was determined to what extent the average score assigned across two, three, four, five, six, seven, eight, and nine assessors matched the average score assigned across ten assessors. This analysis was also conducted twice; once including extreme assessors and once excluding them.

3.3.8 Analysis: Generalization across video episodes

Two analyses were conducted in order to determine to what extent scores assigned to teachers' coaching performance in separate video episodes could be generalized to a universe of intended video episodes. First, a general analysis was conducted that provided an overview of which video episodes provoked varying scores. The results of this analysis do not allow direct conclusions to be drawn with regard to the generalization of scores to a universe, but they do provide information on video episodes that are a threat to the generalizability. For each video episode, the standard deviation of assigned scores across all twelve assessors was determined. When the standard deviation was smaller, the video episodes evidently provoked similar scores; when it was bigger, the video episodes provoked varying scores. Next, a ranking order of video episodes was made, from low standard deviations to high standard deviations. Especially the video episodes low in the ranking order (video episodes with a high standard deviation) were a threat to the generalizability to the universe of video episodes.

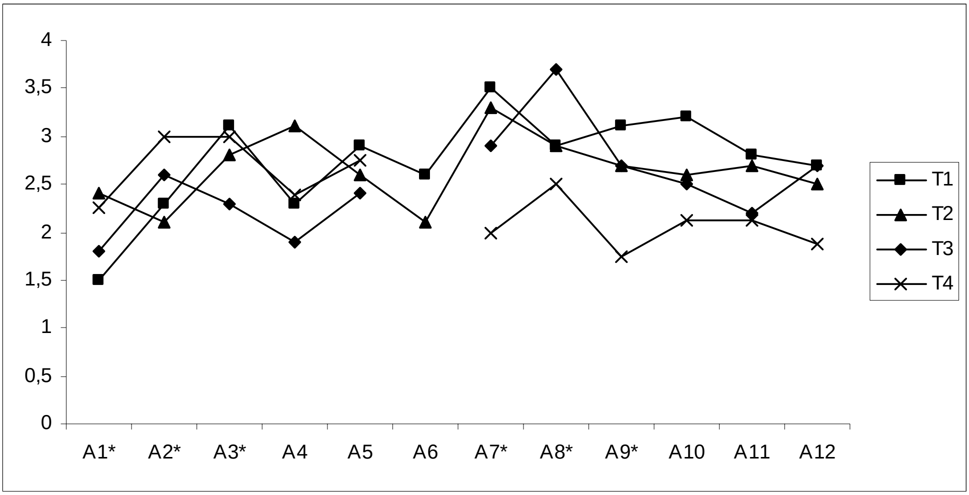
In a second analysis it was determined to what extent a score assigned to a specific video episode matched the scores assigned to other video episodes of the same type. The agreement in assigned scores to the video episodes was used to obtain an indication of the generalizability to the universe of video episodes. In the video portfolios constructed, four types of video episodes were included: video episodes in which the teacher coached in cognitive, meta-cognitive, affective, and collaborative learning activities. The different types of video episodes each formed a separate assessment scale. All video episodes belonging to the same assessment scale were expected to enable measurement of the same construct, and, thus, it should be possible to generalize scores to a universe of video episodes. The better the scores can be generalized, the less video episodes are needed for inclusion in the video portfolio in order to establish an acceptable level of reliability and validity. For each score assigned to a video episode, it was determined to what extent it matched the average remaining score of the assessment scale of which it was part. An average remaining score was the average score assigned to all video episodes that were part of the assessment scale, excluding the video episode for which the correspondence was to be determined. The correspondence between the scores and the average remaining score was expressed in a Gower coefficient.

3.4 Results

3.4.1 Assessors' scoring

Tendencies in scores assigned by assessors

Figure 3.2 presents the average scores assigned by the assessors to the coaching competence of each teacher. Figure 3.2 shows that the lines in the chart are interrupted for teachers three and four. This is because assessor six did not score the video portfolios of teachers three and four. Figure 3.2 shows clearly that the lines in the chart are not parallel to each other. This means that the teachers were not judged equally leniently or severely by the different assessors. The results of the analysis regarding the overall scores are the same. The lines in that chart are not parallel either, which indicates an interaction effect between assessors and teachers.



* These assessors were colleagues of the teachers assessed

Figure 3.2 Average of scores assigned across ten video episodes for twelve assessors for teachers 1, 2, 3, and 4

Based on the findings of these analyses, it appears that mainly colleagues of the teachers judged assigned extreme scores. Figure 3.2 shows that assessor one gave the most severe judgment to teacher one. Assessors two and six assigned the most severe judgment to teacher two; assessor one to teacher three; and assessor nine was the most severe assessor for teacher four. Figure 3.2 also allows the most lenient assessors for each teacher to be determined. Subsequently, it was determined which assessors assigned extreme scores. In 90% of the cases, an extreme score was assigned by an assessor who was a colleague of the teachers assessed. In 60% of the cases, an extreme overall score was assigned by an assessor who was a colleague of the teachers assessed.

Interrater agreement: Frequency

It was first determined for how many cases more than 50% of the assessors assigned exactly the same score to the coaching performance in the video episodes in all four video portfolios. Second, the number of cases was determined for which assessors assigned exactly the same overall score. For teacher one, it was found that more than 50% of the assessors assigned the same score for six of the ten video episodes. For teacher two, this was found for eight of the ten video episodes; for teacher three, for only three of the ten video episodes; and for teacher four, for three out of eight video episodes. These results indicate that the assessors reached more agreement with regard to teachers one and two than for teachers three and four. The frequencies of the overall scores were consistent with the results for the video episodes. Also in assigning overall scores, the assessors reached more agreement with regard to teachers one and two than for teachers three and four.

Interrater agreement: Gower coefficient

Table 3.1 presents the average Gower coefficients across all possible assessor pairs for video episodes and overall scores. The Gower coefficients are presented for each teacher; the ranges of the Gower coefficients found are also presented.

Table 3.1 Gower coefficients for scores assigned to video episodes and overall scores assigned

	Scores assigned to video episodes	Range of Gower coefficients	Overall scores assigned	Range of Gower coefficients
All teachers 38 video episodes/12 assessors 11 overall scores/12 assessors	0.74 0.73*	0.63-0.87 0.56-0.85*	0.80 0.78*	0.61-0.95 0.53-0.95*

Table 3.1 Gower coefficients for scores assigned to video episodes and overall scores assigned (Continued)

	Scores assigned to video episodes	Range of Gower coefficients	Overall scores assigned	Range of Gower coefficients
Teacher 1 10 video episodes/12 assessors 3 overall scores/12 assessors	0.80 0.75*	0.56-0.93 0.33-0.93*	0.79 0.75*	0.33-1.00 0.33-1.00*
Teacher 2 10 video episodes/12 assessors 3 overall scores/12 assessors	0.80 0.78*	0.59-0.92 0.54-0.92*	0.93 0.85*	0.78-1.00 0.56-1.00*
Teacher 3 10 video episodes/11 assessors 3 overall scores/10 assessors	0.71 0.68*	0.52-0.85 0.37-0.90*	0.76 0.68*	0.56-1.00 0.22-1.00*
Teacher 4 8 video episodes/11 assessors 2 overall scores/11 assessors	0.76 0.73*	0.63-0.90 0.57-0.92*	0.82 0.82*	0.67-1.00 0.67-1.00*

* Gower coefficient when extremely lenient and severe assessors were included in the analysis

The Gower coefficient for interrater agreement concerning video episodes was between 0.71 (teacher three) and 0.80 (teachers one and two) when extreme assessors were excluded from the analyses. When extreme assessors were included, the Gower coefficients were somewhat lower (between 0.68 and 0.78). These Gower coefficients indicate that an acceptable level of agreement was reached for the scoring of video episodes. The Gower coefficients for the assignment of overall scores was between 0.76 (teacher three) and 0.93 (teacher two) when extreme assessors were excluded from the analyses. The level of interrater agreement for assignment of overall scores can be regarded as high. When extreme assessors were included in the analyses, the Gower coefficient dropped again (between 0.68 and 0.85), but this can still be considered an acceptable level of agreement.

Generalizability across assessors

The interrater agreement for the average score between two assessors and the average score across ten assessors appeared to be 0.88 to 0.91. These results indicate that the average score based on ten assessors can be estimated quite accurately based on the average score between two assessors. When the extreme assessors were included in the analysis, Gower coefficients were found to be between 0.72 to 0.90 for the average score across two assessors and across twelve assessors. Even when extreme assessors were included, an acceptable to high level of consistency was found for average scores across two and twelve assessors.

3.4.2 Generalization across video episodes

Interrater agreement for specific video episodes

The ranking order of video episodes from low to high standard deviation for scores assigned across assessors was divided into three groups: group one consisted of video episodes for which assessors' scores varied across two scale points on the four-point scale (standard deviation of 0.37-0.49); group two consisted of video episodes for which assessors' scores varied across three scale points (standard deviation of 0.51-0.79); and group three consisted of video episodes for which assessors' scores varied across four scale points (standard deviation of 0.83-0.99). In total, 38 video episodes were judged. Of these, 8 video episodes were in group one, 17 in group two, and 13 in group three. The video episodes that elicited similar scores were in group one, the video episodes that elicited different scores were in group three. The video episodes from group one showed mainly the coaching of teachers one and two in cognitive learning activities. The video episodes from group two showed mainly the coaching of teachers one and two in meta-cognitive learning activities. Video episodes showing teacher four's coaching in collaborative learning activities were also included in this group. The video episodes that elicited different scores from assessors were those of teacher three. Four out of the six video episodes showing coaching in affective learning activities were included in this group.

Agreement on scores assigned to a video episode and the average remaining score

Table 3.2 presents for each video episode the Gower coefficient as an indicator of agreement on the average score across assessors for coaching performance in the specific video episode and the average scores assigned to all other video episodes from the scale to which the specific video episode belongs.

Table 3.2 Gower coefficients for agreement on the average of the scores assigned to a video episode and the average of the scores assigned to the other video episodes of the scale

Video episodes	**Teacher 1**	**Teacher 2**	**Teacher 3**	**Teacher 4**
Cognitive 1	0.81	0.83	-	-
Cognitive 2	0.83	0.72	-	-
Cognitive 3	0.80	0.83	-	-
Cognitive 4	0.78	-	-	-
Meta-cognitive 1	0.82	0.82	0.78	0.70
Meta-cognitive 2	0.85	0.83	0.74	0.80
Meta-cognitive 3	0.89	0.82	0.72	0.77
Meta-cognitive 4	-	0.81	0.64	-
Meta-cognitive 5	-	0.71	-	-
Collaborative 1	-	-	0.66	0.73
Collaborative 2	-	-	0.78	0.80
Collaborative 3	-	-	0.74	0.86
Collaborative 4	-	-	0.66	0.79
Collaborative 5	-	-	-	0.78
Affective 1en 2	0.78	0.67	0.53	-

Table 3.2 shows that, in general, for video episodes pertaining to teachers' coaching in cognitive learning activities, a high level of agreement was found for scores assigned to other video episodes showing coaching in cognitive learning activities. This result indicates that scores assigned to a video episode showing coaching in cognitive learning activities can reasonably be generalized to the universe of video episodes showing coaching in cognitive learning activities. The results regarding the agreement in scores assigned to video episodes concerning coaching in meta-cognitive learning activities show an ambiguous picture. For the video episodes of teachers one and two regarding coaching in meta-cognitive learning activities, a high level of agreement was found. Thus, the scores assigned to these video episodes can reasonably be generalized to the universe of video episodes showing coaching in meta-cognitive learning activities. For the video episodes of teachers three and four, a lower level of agreement was found, which indicates a lower level of generalizability of scores to the universe of video episodes. Furthermore, Table 3.2 shows that the agreement on scores assigned to video episodes concerning coaching in collaborative learning activities is acceptable. As was the case with the video episodes concerning coaching in meta-cognitive learning activities, the scores assigned to these video episodes were less consistent, resulting in a lower level of generalizability to universe of video episodes showing the coaching in collaborative learning activities. The agreement

between scores assigned to video episodes showing affective learning activities is the most problematic. For these video episodes, a low to acceptable level of agreement was found. For video episodes regarding coaching in affective learning activities, it is very difficult to generalize a score to the universe of video episodes showing coaching in affective learning activities.

3.5 Conclusion and discussion

The aim of this study was to examine the extent to which the design principles mentioned in the literature contribute to valid and reliable performance assessments. The specific research questions were, (1) To what extent are assessors capable of scoring teachers' coaching competence in a reliable and valid way based on a video portfolio? and (b) To what extent can scores assigned to the coaching performance in separate video episodes be generalized to the universe of intended video episodes?

Assessors' scoring

The first conclusion that can be drawn is that scoring tendencies occurred in the process of assigning scores. Assessors seemed not capable of scoring the different teachers equally leniently or severely. It is hard to explain why the assessors were not capable of consistent scoring. It might be that it was hard to score consistently, because each teacher coached in a different context or it might be that assessors were influenced by personal biases and preferences for a specific coaching style (Gipps, 1994; Moss, 1994). Furthermore, it appeared that some assessors assign extreme scores in judging their colleagues. This tendency appears in the assignment of scores to teachers' coaching performance in video episodes as well as in the assignment of overall scores. Assessors are extremely lenient as well as extremely severe in assigning scores to their colleagues. The tendency to judge colleagues leniently is addressed in the literature. It is known that assessors who are close to the person to be judged are tempted to be lenient (Aronson, Wilson, & Akert, 2007). However, the results show that assessors judging their colleagues also assign extremely severe scores. There is no clear reason for assessors to do this; maybe personal traits of assessors play a role in this. Furthermore, nothing can be concluded with regard to the validity or appropriateness of the scores assigned by assessors in judging their colleagues. Perhaps these assessors assign more valid scores, because they have more information about the teacher that is relevant to the judgment of the teacher's coaching competence (Schutz & Moss, 2004). It is also possible, however, that in judging their colleagues, assessors are influenced by their biases and expectations concerning the colleagues, despite the highly structured scoring procedure.

A second conclusion that can be drawn is that assessors reached an acceptable level of agreement in the scores assigned, as expressed on the scale showing four levels of performance. An acceptable to high level of agreement was found for the assignment of scores to video episodes in the video portfolios (0.71 to 0.80). For the assignment of overall scores, a high level of agreement was reached in most cases (0.76-0.93). A somewhat lower level of agreement was found when assessors who assigned extreme scores were included in the analyses. However, an acceptable level of agreement was still found (for video episodes, 0.68-0.75, and for overall scores, 0.68-0.85). The difference in agreement between scores assigned to video episodes and overall scores is consistent with results from a previous study (see chapter 2). Furthermore, the assessors indicated in an interview that a single video episode was difficult to score, because it shows only a part of the interaction between teacher and students. In that same interview, assessors pointed out that they acquired a clear view of teachers' coaching competence based on five to six video episodes. A third conclusion is that scores expressed on the four-level performance scale can reasonably be generalized across assessors. The results show that an acceptable level of consistency was reached between the average score assigned across two assessors and across ten assessors (0.88-0.90). When extreme assessors were included in the analyses, the level of consistency was somewhat lower (0.72-0.90). The results implicate that, in practice, it should be feasible to achieve an acceptable level of agreement when two assessors are involved in judging video portfolios. This is an important conclusion, because it is often not possible to involve ten to twelve assessors in an assessment.

Based on these three conclusions, the assumption can be justified that the design principles support the first inference of the validity argument (Kane, 2006). The scoring guide, the performance levels, the scoring procedure, the training, and the composition of the video portfolio generally coincide with reliable scoring by assessors.

Generalization across video episodes

The results show that, in some cases, the scores assigned to a specific video episode can reasonably be generalized to the universe of video episodes, but in other cases the generalization is problematic. Scores assigned to video episodes concerning coaching in cognitive learning activities can reasonably be generalized to the universe of video episodes showing coaching in cognitive learning activities, which indicates that fewer of these video episodes are needed in a video portfolio to establish a valid and reliable assessment. The scores assigned to the video episodes of teachers one and two concerning coaching in meta-cognitive learning activities can reasonably be generalized to the universe of video episodes concerning coaching in meta-cognitive learning activities, but the scores assigned to the video episodes of teachers three and four concerning meta-cognitive learning activities are less generalizable. It is hard to predict why some video episodes can be better generalized than others. Perhaps teachers one and two reacted more consistently in the different video episodes, and teachers three and four showed very different performances. It is also possible that the assessors, somehow, succeeded in scoring the coaching of teachers one and two in a consistent way, and failed to do so for teachers three and four. The level of generalizability of the scores assigned to video episodes concerning coaching in collaborative coaching activities is acceptable for teacher three and high for teacher four. Also in this case, it hard to explain the differences in level of generalizability between the scores assigned to the performances of teachers three and four. The generalizability of the scores assigned to video episodes concerning coaching in affective learning activities appeared to be problematic. A possible explanation for this low level of generalizability is that, in practice, coaching in affective learning activities happens very subtly and is often interrelated with coaching in other learning activities. This makes it difficult for assessors to score the coaching in affective learning activities consistently. In the scoring guide, the coaching in affective learning activities should be defined in more detail, so that assessors have better knowledge of the coaching in affective learning activities at the four different performance levels. Furthermore, the low level of generalizability may be caused by the small number of video episodes included in the video portfolio with regard to coaching in learning attitude.

Only tendencies with regard to the generalizability of scores across video episodes can be described on the basis of the results of this study. No conclusions can be drawn regarding the minimum number of video episodes needed to establish an acceptable level of validity. The standardization of video episodes based on a definition for key situations appeared to go together with predominantly positive effects on generalizability. The agreement on scores assigned to a specific video episode and the average score assigned to other video episodes of the same assessment scale is predominantly acceptable to high; only the agreement on video episodes concerning coaching in affective learning activities is problematic.

Extrapolation to performance outside the assessment context

The tenability of the third inference, addressing extrapolation from the performance shown in the video episodes to performance outside the assessment context, was not investigated in this study. However, some remarks can be made with regard to this inference. The tenability of this inference is likely to be assured by the use of very authentic coaching situations and by establishing variety in the sample of video episodes selected. In putting together a sample of video episodes, we found that it takes a lot of time to collect enough authentic situations representing a variety of coaching situations in which all different learning activities are to be addressed. This was because we were dependent on students' need for support. It is possible that the students were predominantly encountering problems in the performance of cognitive learning activities and needed less support in performing the other three types of learning activities. As a result, there was little choice for the selection of episodes showing the coaching of affective learning, and far more choice for the selection of episodes addressing the other learning activities. In order to determine to what extent the sample of video episodes used in this study is representative of all coaching situations in practice, additional research is needed in the form of a job analysis.

Future research

In this study, it was examined to what extent assessors score teachers' coaching performance in a reliable way. However, in order to get a complete picture of the validity of the assessment procedure, assessors' use of the scoring guide and conceptual framework should also be investigated. This can be done through qualitative analyses, involving the evidence and arguments the assessors use to justify the scores assigned. These analyses may also provide more information about the reasons why assessors judge their colleagues more leniently or severely. In order to be able to draw more decisive conclusions about the minimum number of video episodes needed for a valid assessment, a research design based on a larger number of scored video episodes is needed. When more video episodes are scored, a generalizability study can be done on the scores assigned. These analyses reveal how much variance can be attributed to the different aspects of a performance assessment (assessors, tasks, person, and interaction effects). Furthermore, based on the findings of these analyses, conclusions can be drawn about the number of video episodes needed for a valid assessment.

References

Aronson, E., Wilson, T.D., & Akert, R.M. (2007). *Social psychology* (5th ed.). Amsterdam: Pearson Education Benelux BV.

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.

Borsboom, D., & Mellenbergh, G.J. (2004). The Concept of Validity. *Psychological Review, 11*(4), 1061-1071.

Brennan, R.L. (2000). Performance assessments from the perspective of generalizability theory. *Applied Psychological Measurement, 24*(4), 339-353.

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.

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

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.

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

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.

Dunbar, S.B., Koretz, D.M., & Hoover, H.D. (1991). Quality control in the development and use of performance assessments. *Applied Measurement in Education*, *4*, 289-303.

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.

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.

Kane, M.T. (2004). Certification testing as an illustration of argument-based validation. *Measurement, 2*(3), 135-170.

Kane, M.T. (2006). Validation. In R.L. Brennan (Ed.), *Education Measurement* (4th ed.). Westport: PraegerPublishers.

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.

Linn, R.L., & Burton, E. (1994). Performance-based assessment: Implications of task-specificity. *Educational Measurement: Issues and Practice, 13*(1), 5-15.

Messick, S. (1989). Validity. In R.L. Linn (Ed.), *Educational measurement* (3rd ed.). New York: MacMillan.

Miller, M.D., & Linn, R.L. (2000). Validation of performance assessments. *Applied Psychological Measurement 24*(4), 367-378.

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 inivronments 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.

Peterson, K.D., Stevens, D., & Mack, C. (2001). Presenting complex teaching evaluation data: Advantages of dossier organization techniques over portfolios. *Journal of Personnel Evaluation in Education*, *15*(2), 121-133.

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

Ruiz-Primo, M.A., Baxter, G.P., & Shavelson, R.J. (1993). On the stability of performance assessments. *Journal of Educational Measurement, 30*, 41-53.

Sanders, P.F. (1998). In W.P. van der Brink, en G.J. Mellenbergh (Eds.), *Testleer en testconstructie (Testing and test construction)*. Amsterdam: Boom.

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.

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/.

Shavelson, R.J., Baxter, G.P., & Gao, X. (1993). Sampling variability of performance assessments. *Journal of Educational Measurement, 30*, 215-232.

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

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.

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

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

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.

Chapter 4

The impact of construct-irrelevant variance and construct under-representation in assessing teachers' coaching competence[3]

Abstract

The aim of this study was to investigate the extent to which assessors justify their scores of teachers' coaching competence based on similar evidence and arguments. The evidence used and arguments made by the assessors were investigated with regard to their (ir)relevance and (in)appropriateness. Previous to this study, an authentic teacher-assessment procedure was developed for assessing teachers' coaching competence in the context of senior secondary vocational education (see chapter 2). In this assessment procedure, trained assessors judge 'video portfolios'. A video portfolio consists of video recordings of systematically selected video episodes showing the teachers' coaching performance and context information about the students, the tasks they worked on, etc. In this study, twelve assessors scored four video portfolios. Filled-out score forms containing reported evidence and arguments for assigning a specific score to each video episode were collected and analyzed. Three conclusions were drawn. First, a considerable amount of variation was found in the evidence and arguments reported by the assessors in scoring the same coaching performance, even when assessors assigned the same score to the coaching performance. Second, more variation was found in reported arguments used to justify a score than in reported evidence. Third, assessors were reasonably capable of reporting evidence and arguments that corresponded with the scoring guide and the related conceptual framework for assessing teachers' coaching competence, but tended to focus on different aspects of the conceptual framework.

4.1 Introduction

Much attention is currently given to the design and use of authentic performance assessments in teacher education and for teachers' further professional development.

[3] This chapter has been submitted in adapted form as:
Bakker M., Beijaard, D., Roelofs, E., Tigelaar, D., Sanders, P., & Verloop, N. The impact of construct-irrelevant variance and construct under-representation in assessing teachers' coaching competence.

Typically, in performance assessment, the teacher is asked to perform, produce, or create something over a sufficient duration of time to permit evaluation of either the process or the product of performance, or both. In these types of assessments, the assessment tasks used are open-ended and complex. An important issue in the design and use of performance assessments is how to warrant validity. Validity is a characteristic not so much of the performance-assessment instrument itself, but rather of the way it is used. Messick stated that "validity is an integrated evaluative judgment of the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of inferences and actions based on test scores or other modes of assessment (1989, p. 13)."

A procedure by which an assessment procedure can be evaluated was recently described by Kane (2004, 2006), and summarized using the concept of 'validity arguments'. Kane posited that the validity of an assessment procedure can be evaluated by examining the inferences on which a score is based. Kane distinguishes three interrelated stages in a so-called chain of inferences: scoring performance on assessments tasks, generalizing across assessment tasks towards a universe of tasks, and extrapolating towards the practical domain. In a validity argument, the plausibility of the inferences is evaluated.

This study was focused on the first inference of the validity argument: the evaluation of the quality of teacher performance-assessments scoring. In determining this, interrater agreement or reliability is usually seen as the most important indicator. Accomplishing reliable scores of performance assessments appears to be a serious problem in performance assessments (Gipps, 1994; Moss, 1994). The contexts in which the assessment tasks take place often vary a lot. Furthermore, respondents may react to the assessment tasks in very different ways. It is not easy for assessors to interpret and judge in a consistent way the very different kinds of information that originate from different contexts. Especially selective observation and personal beliefs and views of assessors are threats to the reliable scoring of task performance (Gipps, 1994; Moss, 1994).

In investigating the reliability of performance-assessment scoring, most researchers have only reported the outcomes of the scoring procedure in terms of interrater agreement or reliability. However, interrater agreement statistics lack information about the process of scoring, about the actual use of the scoring rules by raters (Linn, 1994; Messick, 1995; Moss, 1994; Van der Schaaf, Stokking, & Verloop, 2005). Assessors may agree on the scores assigned, but do they also agree on the evidence and arguments that underlie these scores? Do they assign the same scores based on similar evidence and arguments, or based on very different evidence and arguments? Little is known about the evidence and arguments that underlie the scores of individual raters. The aim of this study was to investigate the evidence and arguments that assessors use to justify the scores assigned.

Previous to this study, a performance assessment procedure was developed, aimed at assessing teachers' coaching competence in the context of senior secondary vocational education (see chapter 2). Along with the implementation of competence-based teaching in the Netherlands, coaching has become an important teacher competence. It is expected that teachers who take on a coaching role will contribute to self-regulated and independent learning on the part of the learners, which is one of the central aims of competence-based learning in vocational education (Moerkamp, De Bruijn, Van der Kuip, Onstenk, & Voncken, 2000; Onstenk, 2000). One way to establish a competence-based learning environment is to have teachers coach students who work collaboratively in small groups on complex tasks. In the present study, a video portfolio assessment procedure was used to assess teachers' coaching competence. Based on the work of Frederiksen, Sipusic, Sherin, and Wolfe (1998), the main elements of the video portfolio are video episodes of teachers' coaching performance in the classroom. In order to interpret and judge in a valid way teachers' performance shown in the video episodes, supporting data sources were added that outlined the context in which the coaching took place. The procedures for scoring and judging the video portfolios are outlined in detail in section 3.1.

Four video portfolios of four teachers were constructed and subsequently scored by twelve trained assessors. Data were collected with regard to the reported evidence and arguments underlying an assigned score. The following research questions were answered in this study:

- To what extent do assessors justify their scores assigned to teachers' coaching performance as shown in video episodes using similar evidence and arguments?
- What kind of evidence and arguments do assessors report on score forms?
- To what extent do assessors report evidence and arguments that correspond with the scoring guide and related conceptual framework for assessing competent coaching?

4.2 Threats to validity and reliability

Each assessment is aimed at measuring a specific construct. This specific construct is expected to be embedded in a conceptual framework (Gipps, 1994) that provides a clear and detailed definition of the construct and that makes clear in what way the assessment scores are related to the construct. The conceptual framework is used by assessors during the scoring process. In relation to measuring a specific construct, the literature indicates several threats to a valid scoring process. Table 4.1 provides an overview of these threats. The threats are ordered according to two major threats distinguished by Messick (1995): construct irrelevance and construct under-representation. The distinction between construct irrelevance and construct under-representation can be a useful starting point for investigating the reported evidence and arguments that underlie assessors' scores (see Nijveldt, 2007). In cases of construct irrelevance, assessors base their judgment on evidence and arguments that are not related to the conceptual framework and the construct being assessed, but to other, irrelevant constructs. 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). Schemata are comparable to personal constructs (Kelly, 1995) that are used to organize and interpret information. The use of these (personal) constructs during the scoring can lead to selective observation and to the use of personal beliefs about competent and incompetent performance (Van der Schaaf, Stokking, & Verloop, 2005). The findings of recent studies focused on construct irrelevance confirmed that assessors were applying irrelevant, personal constructs (Baume, York, & Coffey, 2004; Frederiksen, Sipusic, Sherin, & Wolfe, 1998; Moss & Schutz, 2004; Van der Schaaf, Stokking, & Verloop, 2005). Other research findings showed that assessors were reasonably capable of applying criteria from the conceptual framework that they were supposed to use during scoring (Heller, Sheingold, Myford, 1998; Nijveldt 2007). In cases of construct under-representation, assessors fail to capture critical evidence and arguments related to the construct being assessed. Construct under-representation can be the result of different kinds of scoring processes. As shown in Table 4.1, construct under-representation can be caused by an inappropriate emphasis on particular evidence and arguments (threat 2). As part of this threat, selective observation is a well-known phenomenon; assessors select just a part of the relevant evidence and/or take just a part of the relevant evidence and arguments into account in assigning a particular score, so that critical aspects of the construct are missed. Construct under-representation can also be caused by making interpretations and judgments that are too analytic (threat 3). When assessors score performance too analytically, they focus on too-small aspects of the performance and do not capture the richness of the whole performance. Furthermore, construct under-representation can be caused by scoring too holistically (threat 4). When assessors score the performance too holistically, they focus only on the general aspects of the performance, so that they miss relevant and more detailed aspects. Especially when assessors focus on the performance as a whole, there is a risk that they will make inferences and judgments that are not entirely based on relevant evidence, but on their personal assumptions and biases (Klein & Stecher, 1998). Finally, construct under-representation can occur when assessors do not apply the conceptual framework and/or the scoring procedure consistently (Crooks, Kane, & Cohen, 1996) (threat 5). Although the above-mentioned threats have been recognized, they have not yet been investigated in-depth.

Table 4.1 Overview of threats to the validity of assessors' scoring processes

Construct irrelevance	**Construct under- representation**
1. Assessors apply extraneous criteria which are not related to the construct being assessed	2. Assessors place inappropriate emphasis on particular evidence and arguments 3. Assessors make interpretations and judgments that are too analytic 4. Assessors make interpretations and judgments that are too holistic 5. Assessors do not apply the conceptual framework consistently

In order to investigate reported evidence and arguments, we started by investigating *what* evidence and arguments assessors identify, select, and use to justify assigned scores. Applying extraneous criteria and placing inappropriate emphasis on particular evidence and arguments (threats 1 and 2) can play a role in these processes, and were investigated in this study. Making interpretations that are too analytic or too holistic, and applying the scoring rules and related conceptual framework in an inconsistent way (threats 3, 4, and 5) are relevant in other parts of the scoring process, like in combining evidence and arguments to make an overall judgment and in assigning scores to teachers' coaching performances. These processes are also relevant parts of the scoring process, but were not the topic of this research.

In order to minimize the occurrence of construct-irrelevant variance and construct under-representation, several measures have been proposed in the literature. The most important measure to reduce these threats is to train assessors in applying the scoring rules related to the relevant constructs from the conceptual framework (Day & Sulsky, 1995; Stamoulis & Hauenstein, 1993; Woerh & Huttcuff, 1994). Other measures pertain to the quality and transparency of the scoring rules and conceptual framework used during the assessment (Crooks, Kane, & Cohen, 1996; Gipps, 1994; Frederiksen, Sipusic, Sherin, & Wolfe, 1998; Kane, 2006; Linn, Baker, & Dunbar, 1991). These measures are summarized in Table 4.2. In section 4.3.1, it is described in detail how these measures were elaborated in the design of the assessment used in this research.

Table 4.2 Overview of measures for reducing the impact of construct irrelevance and construct under-representation in authentic assessments

Construct irrelevance	**Construct under- representation**
- Use a conceptual framework that includes only relevant aspects of the construct - Train the assessors in applying the scoring rules and related conceptual framework in a systematic and consistent way	- Use a conceptual framework that includes only relevant aspects of the construct - Use scoring rules that are systematic and transparent - Train the assessors in applying the scoring rules related to the conceptual framework in a systematic and consistent way - Train assessors in avoiding rating errors

4.3 Method

4.3.1 Design of the assessment procedure

Video portfolios

In the present study, assessors judged teachers' coaching competence based on a video portfolio. The video portfolios consisted of a mix of sources of evidence that were expected to provide the assessors with a complete picture of the teachers' coaching competence. The main sources of evidence consist of video episodes that represent coaching performance. For this, the teachers were filmed on-the-job during coaching sessions with a group of students. The video recordings represent performance in an authentic context. In order to be able to score and judge the teachers' coaching performance in the video-recorded episodes in a valid way, information about the context was added: interviews with the teachers about the decisions underlying their actions; interviews with students about the perceived impact of teachers' actions on their work; information about students' backgrounds; information about the learning tasks students worked on during a video episode; information about students' progress in completing the tasks; and information about the teachers' backgrounds. The assessors were expected to examine all these sources while assessing a video portfolio. In addition to these sources of evidence, information was added to the video portfolios about the educational materials students use during the video episodes and students' products that are discussed during video episodes. The assessors were expected to use these sources of evidence in assessing a video portfolio when they felt a need for this extra information in order to gain a better understanding of the coaching situation.

Scoring guide based on a conceptual framework for coaching

In order to reduce the impact of construct-irrelevant variance and construct under-representation in assessors' scoring processes, a scoring guide related to a conceptual framework for coaching was constructed. The main purpose of the scoring guide was to ensure that assessors would pay attention to the characteristics of competent coaching, and in so far as possible to prevent them from scoring and judging video portfolios according to their own personal criteria. The development of the guide and the related conceptual framework was based on a literature study in the field of supporting self-regulated learning and observations of coaching situations in practice.

In the scoring guide, coaching was defined 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 this learning are asking questions and providing feedback on learning activities conducted by students. These coaching interventions were expected to be used to stimulate and support four types of learning activities: cognitive, meta-cognitive, and affective learning activities (Shuell, 1993; Vermunt & Verloop, 1999; Winne & Hadwin, 1998), and activities related to collaborative learning (Johnson & Johnson, 1994; Perry, 1998; Perry, Phillips, & Dowler, 2004; Slavin, 1990). Cognitive learning activities concern activities students use to process subject matter, resulting in 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 activities are thinking activities 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). Collaborative learning activities concern activities with regard to communication, coordination, and realisation of a positive group climate (Johnson & Johnson, 1994; Slavin, 1990).

The scoring guide was expected to assist assessors in scoring teachers' coaching performance in a systematic and consistent way. First, concrete examples of coaching interventions were included in the scoring guide, so that assessors were better capable of recognizing relevant coaching interventions. When they know better what to judge, assessors are less inclined to apply their personal constructs and criteria (Frederiksen, Sipusic, Sherin, & Wolfe, 1998). Second, a criterion for competent coaching and several performance levels were elaborated. 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 the 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. 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, coaching was considered competent when teachers used coaching interventions that provided students with opportunities to improve their learning activities. In this study, competent coaching was defined as constructive coaching. In constructive coaching, the teacher provides just enough support so that the students can take the step to a higher level in undertaking learning activities, which they couldn't have taken on their own (Vygotsky, 1978). As improvements in performing a learning activity increases, the support of the teacher decreases, until the student can perform the learning activity by him/herself; 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 performance 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). Four levels of performance were formulated based on the criterion of constructive coaching. For each level, illustrative level descriptors were made. The descriptors were expected to assist assessors in making relevant considerations and in deciding which performance level matches the observed coaching performance. The performance levels are presented in Table 2.2 in chapter 2.

Scoring procedure

The assessors were expected to score the video portfolios according to a detailed scoring procedure. In this procedure, the assessors were asked to start by collecting specific evidence pertaining to teachers' questions and feedback that did or did not provide an opportunity for students to improve their performance of learning activities. Subsequently, the assessors were to use the specific evidence to build a judgment concerning the performance across the whole episode (in this case, whether the teacher did/did not contribute to students' growth). Furthermore, the assessors were expected to form an overall judgment about the teachers' coaching competence based on their performance across the video episodes. The steps in the scoring procedure are presented in Table 2.5 in chapter 2 and the score forms used in Appendix 2 and 3. The assessors were urged to follow the steps of the scoring procedure in detail. In Table 2.5 and on the score forms presented in Appendix 2 and 3, instructions are included for scoring to what degree teachers' coaching performance was practice-oriented. However, in this study, assessors were asked to score teachers' coaching performance only for constructive coaching. This decision was based in the results of study 1, which showed that practice-oriented coaching could not be scored in a valid way based on the video portfolios constructed.

The scoring procedure was elaborated along with the measures to reduce the impact of construct under-representation described by Moss, Schutz, and Collins (1998) and Schutz and Moss (2004). The first measure is that assessors should use all available evidence in making a judgment. For that reason, the assessors were urged in the instructions to consider all available evidence and to check afterwards whether they had based the assigned score on all available evidence. The second measure is that assessors should actively seek counter-evidence in order to reduce the impact of construct under-representation. In the scoring procedure, the assessors were urged to search for coaching interventions demonstrated by the teacher that did provide opportunities for students as well as interventions that did not. The third measure is that 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. In order to give assessors an opportunity to exchange interpretations and judgments with another assessor, a discussion was included in the scoring procedure (step 4).

Assessor training

Assessor training is a prerequisite for accurate ratings and to reduce the impact of construct-irrelevant variance and construct under-representation in performance assessments (Stamoulis & Hauenstein, 1993; Day & Sulsky, 1995; Uhlenbeck, 2002; Woerh & Huttcuff, 1994). For that reason, assessor training was set up to prepare the assessors for scoring and judging video portfolios. Four training sessions were developed that were aimed at enabling assessors to use the designed conceptual framework and the scoring method in a systematic and consistent way.

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 in the following:

- 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 coaching;
- 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, assessors were corrected when they deviated from the scoring procedure. Another important aim of the training was to make assessors aware of rating errors and to have them immediately correct those errors in case they occur. Special attention was given to errors concerning an inappropriate emphasis on specific evidence or arguments, selective observation, inconsistencies in assessors' scoring, halo-effect, horn-effect, and central tendency (Aronson, Wilson, & Akert, 2007).

4.3.2 Materials

The researchers constructed video portfolios of four teachers. The four teachers involved (three male and one female) worked as coaches in a school for senior secondary vocational education, in the building technology section. The teachers had one to two years of experience in coaching students.

4.3.3 Participants

The video portfolios were scored by twelve assessors, i.e., teachers from the same discipline as the teachers to be judged and who had an equal amount of experience in coaching students. Six of the twelve assessors worked at the same school as the teachers recorded in the video portfolios. The other six assessors were from another school.

4.3.4 Data collection

After the four training sessions, the assessors independently scored the four video portfolios. Each video portfolio contained ten video episodes, except for one video portfolio that contained only eight video episodes. The video episodes in a video portfolio cover the range of learning activities to be induced by the coaching, as elaborated in the conceptual framework. The assessors started by scoring the video episodes in the portfolio. Using score forms, the assessors reported which coaching interventions did and which did not give students an opportunity to improve their conducting of a specific learning activity (step 1 of the scoring procedure). Based on the evidence gathered, assessors assigned a score to the coaching performance shown in the complete video episode. In addition, they wrote a summary report on the score form in which they justified the score assigned (step 2 of the scoring procedure). After having scored the separate video episodes, assessors assigned an overall score to the coaching performance across video episodes and wrote a summary report to justify the score assigned (step 3 of the scoring procedure). The summary reports that were written to justify the overall scores were so concise that they did not provide enough information for this study; these summaries were left out of the analysis.

4.3.5 Analysis

The reported coaching interventions in all video episodes and the summary reports from the score forms were used for analysis. Before the analysis took place, score forms were selected. In total, 38 video episodes were scored by twelve assessors. Ten episodes from each of the video portfolios were used for scoring by the assessors; the video portfolio of teacher 4 was an exception. In the latter case, eight episodes were scored. For each video episode the assessors scored, they filled out a score form. In total, 420 score forms were available for analysis. Score forms were selected based on the following procedure. An important criterion for selection was that score forms were included from video episodes for which assessors had reached a high level of agreement on the scores assigned, as well from video episodes for which assessors had reached a low level of agreement on the scores assigned. In that way, we aimed to get more insight into processes that play a role when assessors do and do not reach agreement on assigned scores. The standard deviation of scores assigned across the 12 assessors was used as a standard for agreement with regard to scores assigned to teachers' coaching performance in separate video episodes. When the standard deviation was large, there was less agreement between assessors with regard to assigned scores, and vice versa. Video episodes were ranked based on the standard deviation of scores assigned across the 12 assessors. The six video episodes with the highest standard deviation and the six video episodes with the lowest standard deviation were selected. In total, 126 score forms were selected and analyzed in this study.

Analysis 1: Variation in evidence and arguments reported by assessors

Evidence and arguments were investigated in order to determine to what extent assessors reported corresponding evidence and arguments. In Atlas/ti, codes were assigned to evidence and arguments based on content. During coding, it was found that assessors differed greatly in the amount of evidence they reported. Some assessors reported detailed lists of evidence; others reported only what they believed to be the most important evidence. Whether assessors reported a string of evidence or just one or two interventions from that string, the same content-code was assigned. Table 4.3 presents an example of two score forms filled out by two different assessors. The same content-code (spring bolts) was assigned to the bold-printed strings of evidence in Table 4.3.

Table 4.3 Examples of filled-in score forms

Score form: Assessor 1		Score form: Assessor 2	
Evidence:		Evidence:	
- Teacher: in the overview, sand to fill up….	-	- Do we need sand?	3
- Do we need sand?	2	- Does that fit in the category 'groundwork'?	3
- Does the sand belong to the category 'groundwork' or 'street work'?	3	- Does the sand belong to the category 'groundwork' or 'streetwork'?	3
- So, sand belongs to groundwork? We agreed that we would cluster the activities according to categories	2/3	- What is missing in the category groundwork?	3
- When do we work with sand?	3	- How do we attach the boards?	3
- Teacher explains the differences between groundwork and street work	-	- **What are spring bolts?**	3
- The apron is almost complete, what is missing here? (teacher asks Pete, but John answers; the teacher asks Pete another question)	2/3	- **What do spring bolts look like?**	3
- How do we attach the boards?	3	- Is there a purlin along the boards?	3
- How do we attach the sole?	3		
- **What are spring bolts?**	3/4		
- **What do spring bolts look like?**	3/4		
- **Is it important to know what spring bolts look like?**	3		
- **Gives an example of what could happen in practice; you may receive an order for spring bolts, then it is convenient to know what they look like.**	4		
- Is there a purlin along the boards?	3/4		

Table 4.3 Examples of filled-in score forms (Continued)

Summary report: This coaching session can clearly be divided in three parts: (1) ground and street activities, (2) attaching the apron, and (3) attaching the purlin. The teacher asks the right questions. And after a sequence of questions, he provides the students with a short explanation. He relates the domain-specific knowledge to relevant situations in practice. I think that the students can certainly learn from these interventions. The judgment will be a 3 or 4. The reason for assigning a 3 instead of a 4 is that the teacher provides a lot of theory. I don't think that students who do not take notes will remember what the teacher aims to teach them.	Summary report: This teacher has good coaching sessions, and in this coaching session he uses the right questions to urge students to comprehend and apply the domain-specific knowledge in the right way. He asks the questions in such a way that the students are steered towards the correct approach. The teacher could have gone on to ask questions on domain-specific knowledge in a broader sense.
Judgment: 3	Judgment: 3

Analysis 2: Types of evidence and arguments

For the second analysis, the nature and content of the reported evidence and arguments were coded. The reported evidence and arguments were coded in Atlas/ti, using the codebook described in Appendix 4. The evidence and arguments were coded in four broad categories. The first category pertained to the type of statement that assessors reported. According to the Associated Systems Theory (Carlston, 1992; 1994), and confirmed by the research of Van der Schaaf, Stokking, and Verloop (2005), assessors use evidence and arguments that differ in level of abstraction; assessors use concrete observations as well as abstract inferences to justify an assigned score. For that reason, evidence and arguments in this study were coded for level of abstraction. Not only abstract inferences were found in the data, but also abstract inferences that contained a judgment. For the inferences that contained a judgment, a code 'judgment' was added to the codebook. Evidence or arguments were coded as 'citation' when assessors reported concrete interventions or concrete statements from the video recording (low level of abstraction). Evidence or arguments were coded as 'inference' when assessors reported an interpretation in their own words of what happened in the video recording (high level of abstraction). Evidence or arguments were coded as 'judgment' when assessors made statements in terms of 'good' or 'bad' (high level of abstraction). An example of coded evidence is presented in Appendix 5, and an example of coded arguments is given in Appendix 6. The second coding category referred to the valence of the evidence and arguments in terms of positive, negative, or neutral. In the scoring guide, the assessors were urged to look for positive as well as negative evidence. The evidence and arguments were coded for valence in order to get an indication of the proportion of positive, negative, and neutral evidence. The proportion provides information on assessors' tendency to focus more on positive or on negative evidence or arguments. The third category pertained to the aspects of competent coaching. In the scoring guide, a definition of competent teaching by Roelofs and Sanders (2007) was used to define a criterion for competent coaching. This definition was also used to distinguish the different aspects. Competent teaching was 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 (positive consequences for students), all according to accepted professional standards. This definition includes several aspects: deliberate and appropriate decisions; teachers' actions (behavior); consequences for the students; and the complex, professional context. The evidence and arguments that were reported by assessors in this study were coded into one of these aspects: the context (or coach situation), teachers' behavior, or consequences for students. No reported evidence or argument was found to be related to teachers' decisions. The fourth coding category pertained to the learning activity the evidence or argument was related to. In the scoring guide, assessors were urged to judge the function of coaching for a specific learning activity. The coding in this category provides insight into whether assessors were capable of noting evidence and giving arguments related to the specific learning activities they were supposed to judge. As shown in Appendix 6, not all arguments related explicitly to a specific learning activity. In that case, no code for fostering a learning activity was assigned. Furthermore, as shown in Appendix 4, the codes in the upper half of the four broad categories are codes for reported evidence and arguments that are consistent with the conceptual framework for competent coaching, and the codes in the lower half are codes for reported evidence and arguments that are not consistent with the conceptual framework and are thus irrelevant to teachers' coaching competence.

The interrater agreement (Cohen's $\varkappa$) was determined between the coding of two raters. Score forms (n=12) were coded independently by the author of this dissertation and another researcher who is doing research in the same domain. The

Cohen's Kappa for the total codebook was 0.96. In Table 4.4, the Cohen's Kappa's are presented for each category in the codebook.

Table 4.4 Cohen's ϰ for all categories in the codebook

Category	Evidence	Arguments
Type of statements	0.67	0.80
Valence	1.00	0.96
Aspect of competent coaching	1.00	0.72
Fostered learning activity	1.00	1.00
(in)consistent with the conceptual framework	1.00	0.98

4.4 Results

Results with regard to variation in evidence and arguments reported by assessors

Two frequency tables of content codes were generated. Table 4.5 presents the frequencies of reported evidence and arguments for individual video episodes that are unshared and shared by 2-4, 5-8, and 9-12 assessors. The data come from the score forms of video episodes for which the highest level of agreement was found with regard to scores assigned to teachers' coaching performance. Table 4.6 presents similar frequencies, but pertains to video episodes for which the lowest level of agreement was found. The frequency tables reveal how many assessors reported the same piece of evidence or arguments on their score forms. As shown in Table 4.5, in the scoring of video episode 1, out of the 19 pieces of evidence, 13 (68%) were reported by one assessor, 1 (5%) was reported by 2-4 assessors, 3 (16%) were reported by 5-8 assessors, and 2 (11%) were reported by 9-12 assessors.

Table 4.5 Frequencies with regard to video episodes for which assessors agreed most on the scores assigned to teachers' coaching performance

Video episode		Citations reported by 1 assessor	Similar citations reported by 2-4 assessors	Similar citations reported by 5-8 assessors	Similar citations reported by 9-12 assessors	Total number of citations	Number of assessors
1	Evid.	13 (68%)	1 (5%)	3 (16%)	2 (11%)	19 (100%)	11
	Arg.	26 (93%)	2 (7%)	-	-	28 (100%)	11
2	Evid.	9 (45%)	4 (20%)	4 (20%)	3 (15%)	20 (100%)	11
	Arg.	19 (86%)	3 (14%)	-	-	21 (100%)	11
3	Evid.	14 (56%)	4 (16%)	6 (24%)	1 (4%)	25 (100%)	11
	Arg.	26 (90%)	3 (10%)	-	-	29 (100%)	11
4	Evid.	10 (59%)	7 (41%)	-	-	17 (100%)	9
	Arg.	10 (83%)	2 (17%)	-	-	12 (100%)	9
5	Evid.	20 (49%)	11 (27%)	10 (24%)	-	41 (100%)	11
	Arg.	18 (82%)	4 (18%)	-	-	22 (100%)	11
6	Evid.	19 (61%)	6 (19%)	2 (6%)	4 (13%)	31 (100%)	12
	Arg.	24 (86%)	3 (11%)	1 (4%)	-	28 (100%)	12
Total		**208**	**50**	**26**	**10**	**293**	

First, Table 4.5 reveals that evidence and arguments reported by one assessor occur by far the most frequently. Second, Table 4.5 shows that the variation in arguments reported by assessors is higher than the variation in evidence reported by assessors. The proportion of arguments reported by one assessor is between 82% and 93%.

Table 4.6 Frequencies with regard to video episodes for which assessors agreed least on the scores assigned to teachers' coaching performance

Video episode		Citations reported by 1 assessor	Similar citations reported by 2-4 assessors	Similar citations reported by 5-8 assessors	Similar citations reported by 9-12 assessors	Total number of citations	Number of assessors
1	Evid.	10 (50%)	5 (25%)	3 (15%)	2 (10%)	20 (100%)	10
	Arg.	19 (90%	2 (10%)	-	-	21 (100%)	10
2	Evid.	8 (73%)	-	3 (27%)	-	11 (100%)	10
	Arg.	17 (100%)	-	-	-	17 (100%)	10
3	Evid.	5 (38%)	6 (46%)	2 (15%)	-	13 (100%)	11
	Arg.	7 (58%)	5 (42%)	-	-	12 (100%)	11
4	Evid.	10 (53%)	4 (21%)	3 (16%)	2 (11%)	19 (100%)	10
	Arg.	18 (90%)	2 (10%)	-	-	20 (100%)	10
5	Evid.	31 (74%)	10 (24%)	1 (2%)	-	42 (100%)	9
	Arg.	18 (100%)	-	-	-	18 (100%)	9
6	Evid.	34 (64%)	14 (26%)	5 (9%)	-	53 (100%)	11
	Arg.	18 (86%)	3 (14%)	-	-	21 (100%)	11
Total		**195**	**51**	**17**	**2**	**267**	

Table 4.6 shows that there is little more variation in reported evidence and arguments for the video episodes for which assessors agreed least on the scores assigned to teachers' coaching performance. For these video episodes, a higher percentage of evidence and arguments was found that was reported by one assessor. Furthermore, the total number of similar pieces of evidence and arguments reported by 5-8 and 9-12 assessors is lower than the total number of similar pieces of evidence and arguments reported by 5-8 and 9-12 assessors for video episodes from Table 4.5. Similar to video episodes with a high level of agreement (Table 4.5), more variation was found in reported arguments than in reported evidence. The proportion of arguments reported by one assessor varies between 58% and 100% for the different video episodes.

Results with regard to types of evidence and arguments

Table 4.7 shows the frequencies of the different types of statements made by assessors. Furthermore, the frequencies for valence, aspect of competent coaching, and fostered learning activity are shown in Tables 4.8, 4.9, and 4.10.

Table 4.7 Frequencies of types of statements

	Frequencies of citations	**Frequencies of inferences**	**Frequencies of judgments**	**Total**
Evidence	866 (78%)	195 (17%)	54 (5%)	1115 (100%)
Arguments	8 (2%)	120 (32%)	244 (66%)	372 (100%)

As shown in Table 4.7, assessors used mainly concrete statements as evidence, and mainly abstract judgments in the summary reports in which they justified the score they assigned.

Table 4.8 Frequencies with regard to valence

	Positive	**Negative**	**Neutral**	**Total**
Evidence	431 (39%)	184 (16%)	500 (45%)	1115 (100%)
Arguments	122 (32,5%)	128 (35%)	122 (32,5)	372 (100%)

Table 4.8 shows that assessors reported more positive evidence than negative evidence. In the summary reports, however, assessors reported approximately as many positive arguments as negative arguments.

Table 4.9 Frequencies with regard to perspective on coaching

	Perspective on coaching	**Codes**	**Frequency of citations**	**Frequency of inferences**	**Frequency of judgments**	**Total**
Evidence	Coaching situation	Students' problem	-	9 (0.80%)	-	
		Groups' problem	-	8 (0.70%)	-	
		Content of the coaching session	-	7 (0.60%)	-	
		Aim of the teacher	-	2 (0.20%)	-	
		Context factors that influence the coaching	-	3 (0.30%)	1 (0.09%)	
		Learning climate	-	2 (0,20%)	-	
		Total	**-**	**31 (2,81%)**	**1 (0.09%)**	**32 (3%)**
	Teachers' behavior	Asking questions	649 (58.50%)	23 (2%)	11 (1%)	
		Providing feedback	186 (17%)	75 (7%)	3 (0.30%)	
		Questions and feedback	-	-	1 (0.09%)	
		Other teacher behavior	-	32 (3%)	5 (0.50%)	
		Missed opportunities	6 (0.50%)	16 (1.50%)	-	
		Interventions are (not) appropriate	-	-	25 (2%)	
		Interventions to direct the discussion	-	4 (0.40%)	-	
		Teachers' style	-	2 (0.20%)	2 (0.20%)	
		Teachers' personal traits	-	-	-	
		Total	**841 (76%)**	**152 (14%)**	**47 (4%)**	**1040 (94%)**
	Consequences for students	Students' reactions to the interventions of the teacher	23 (2%)	8 (0.70%)	-	
		Question to the teacher	1 (0.09%)	-	-	
		Reaction to other students	1 (0.09%)	5 (0.50%)	-	
		Total	**25 (2%)**	**13 (1%)**	**-**	**38 (3%)**

Table 4.9 Frequencies with regard to perspective on coaching (Continued)

	Perspective on coaching	**Codes**	**Frequency of citations**	**Frequency of inferences**	**Frequency of judgments**	**Total**
Arguments	Coaching situation	Students' problem	-	4 (1%)	-	
		Groups' problem	-	8 (2%)	-	
		Content of the coaching session	-	17 (5%)	-	
		Aim of the teacher	-	4 (1%)	-	
		Context factors that influence the coaching	-	11 (3%)	-	
		Learning climate	-	2 (0.60%)	-	
		Total	**-**	**46 (14%)**	**-**	**46 (14%)**
	Teachers' behavior	Asking questions	-	10 (3%)	12 (4%)	
		Providing feedback	-	17 (5%)	1 (3%)	
		Questions and feedback	-	5 (1.50%)	23 (7%)	
		Other teacher behavior	-	18 (5.50%)	11 (3%)	
		Missed opportunities	1 (0.30%)	1 (0.30%)	16 (5%)	
		Interventions are (not) appropriate	-	-	79 (24%)	
		Interventions to direct the discussion	-	5 (1.5%)	2 (0.60%)	
		Teachers' style	-	2 (0.60%)	4 (1%)	
		Teachers' personal traits	-	2 (0.60%)	-	
		Total	**1 (0.30%)**	**60 (18%)**	**157 (48%)**	**227 (66%)**
	Consequences for students	Students' learning	-	-	3 (0.90%)	
		Students' thinking	-	4 (1%)	9 (2.50%)	
		Students' understanding	-	4 (1%)	9 (2.50%)	
		Students' growth	-	-	38 (11%)	
		Students' awareness	-	-	3 (0.90%)	
		Total	**-**	**8 (2%)**	**62 (19%)**	**70 (21%)**

Table 4.9 shows that assessors for the most part reported the concrete teacher interventions ‘asking questions’ and ‘providing feedback’ as evidence (76% of all reported evidence). In addition, assessors also used inferences about teacher behavior (14% of all reported evidence). In the summary reports, assessors used mainly inferences and judgments. The inferences in the summary were related to the coaching situation (14% of all arguments) and to teachers’ behavior (18% of all arguments). Inferences with regard to the coaching situation were often used by assessors to start a summary report, and concerned a description of the content of the coaching situation and a description of factors that, in their opinion, had influenced the coaching of the teacher. The inferences with regard to teachers’ behavior concerned mainly providing feedback and ‘other teacher behavior’. The latter category contained arguments that were not explicitly related to teacher interventions, like questions and feedback, but concerned teacher actions such as the teacher checks…., the teacher listens…., the teacher refers to…, the teacher lists…, the teacher directs…., and the teacher takes action. The judgments in the summary were related to teacher behavior (48% of all arguments) and to the consequences of teachers’ behavior for the students (19% of all arguments). Assessors’ judgments mainly pertained to the appropriateness of teachers’ interventions, the quality of the questions and feedback used by the teacher, and the opportunities offered for students’ growth. As Table 4.9 shows, most of the reported evidence and arguments is consistent with the conceptual framework for competent coaching. Only 1% of the evidence and 4.5% of the arguments (codes ‘learning climate’, ‘interventions to direct the discussion’, ‘teachers’ style’, and ‘teachers’ personal traits’) are not consistent with the conceptual framework.

Table 4.10 shows the characteristics of the evidence reported by assessors according to the learning activity fostered. In this table, twelve video episodes are listed in the columns. The first six video episodes are video episodes for which assessors reached a high level of agreement with regard to the scores assigned. Video episodes 7 to 12 are video episodes for which a low level of agreement was reached. For each video episode, assessors were supposed to judge the coaching in a specific learning activity. This specific learning activity is also indicated in the columns of the table. In the rows of Table 4.10, all possible learning activities are listed. In the analysis, all evidence was coded in the category ‘learning activity fostered’. As shown in Table 4.10, for video episode 1, 108 pieces of the reported evidence referred to coaching of comprehending and using relevant subject matter, 1 piece of evidence referred to coaching of motivation and dedication, and 1 to coaching of contribution to the group process and product. Video episode 1 was expected to be judged on comprehending and using relevant subject matter. This means that 2 of the 110 pieces of evidence (2%) can be regarded as irrelevant evidence.

Table 4.10 Frequencies of evidence with regard to fostered learning activity

Fostered learning activity	Video episode 1 judged on comprehending and using relevant subject	Video episode 2 judged on comprehending and using relevant subject	Video episode 3 judged on comprehending and using relevant subject	Video episode 4 judged on group climate	Video episode 5 judged on planning	Video episode 6 judged on comprehending and using relevant subject matter
Coaching of searching and organizing relevant information	-	-	-	-	6 (5%)	23 (24%)
Coaching of comprehending and using relevant subject matter	108 (98%)	58 (88%)	88 (92%)	-	-	63 (65%)
Coaching of planning	-	-	1 (1%)	-	112 (87%)	-
Coaching of monitoring	-	-	-	-	-	3 (3%)
Coaching of adjusting	-	-	-	-	3 (2%)	-
Coaching of motivation and dedication	1 (1%)	7 (12%)	7 (7%)	-	-	8 (8%)
Coaching of communication	-	-	-	11 (31%)	-	-
Coaching of contribution to the group process and product	1 (1%)	-	-	21 (58%)	8 (6%)	-
Coaching of group climate	-	-	-	4 (11%)	-	-
Group dynamics	-	-	-	-	-	-
Total	110 (100%)	66 (100%)	96 (100%)	36 (100%)	129 (100%)	97 (100%)

Table 4.10 Frequencies of evidence with regard to fostered learning activity (Continued)

Fostered learning activity	Video episode 7 judged on contribution to the group process	Video episode 8 judged on contribution to the group process and product	Video episode 9 judged on motivation and dedication	Video episode 10 judged on adjusting	Video episode 11 judged on motivation and dedication	Video episode 12 judged on monitoring
Coaching of searching and organizing relevant information	-	-	-	-	5 (5%)	-
Coaching of comprehending and using relevant subject matter	-	-	-	-	-	-
Coaching of planning	1 (1%)	-	2 (4%)	-	70 (69%)	-
Coaching of monitoring	-	1 (3%)	-	5 (9%)	-	89 (62%)
Coaching of adjusting	-	-	-	53 (91%)	3 (3%)	-
Coaching of motivation and dedication	-	1 (3%)	37 (71%)	-	17 (17%)	43 (30%)
Coaching of communication	10 (10%)	-	-	-	-	-
Coaching of contribution to the group process and product	85 (84%)	34 (91%)	13 (25%)	-	6 (6%)	11 (7%)
Coaching of group climate	5 (5%)	-	-	-	-	1 (1%)
Group dynamics	-	1 (3%)	-	-	-	-
Total	101 (100%)	37 (100%)	52 (100%)	58 (100%)	101 (100%)	144 (100%)

Table 4.10 shows that construct-irrelevant evidence was reported during the scoring of all video episodes analyzed. Slightly less irrelevant evidence was reported during the scoring of video episodes 1 to 6 than during the scoring of video episodes 7 to 12. The same analysis was done for reported arguments. The results are comparable to the

results presented in Table 4.10. Small differences were found in construct-irrelevant arguments reported during the scoring of episodes 1 to 6 compared with the scoring of episodes 7 to 12. For the arguments, however, fewer construct-irrelevant arguments were found in the scoring of episodes 7 to 12 than in the scoring of episodes 1 to 6. Furthermore, it was found that in the summary reports, assessors referred less to specific learning activities.

4.5 Conclusion and discussion

The aim of the study was to investigate the evidence and arguments that assessors used to justify the scores they assigned. A video portfolio assessment procedure was developed; video portfolios of four teachers were constructed and subsequently scored by twelve trained assessors. Score forms were collected, and quantitative as well as qualitative analyses were carried out. We investigated the extent to which assessors justified the scores assigned to teachers' coaching performance shown in a video episode based on similar evidence and arguments. Furthermore, we investigated the kinds of evidence and arguments assessors reported on score forms, and the extent to which the reported evidence and arguments corresponded with the scoring guide and thus with the conceptual framework for competent coaching used.

With regard to the first research question, it can be concluded that slightly more variation was found in reported evidence and arguments for the video episodes for which assessors agreed the least on scores assigned to teachers' coaching performance than in the evidence and arguments for video episodes for which assessors agreed the most on assigned scores. For all video episodes, however, a considerable amount of evidence and arguments was reported by only one assessor. Even when assessors assigned the same score to the coaching performance in a video episode, they based their scores on different evidence and argument. This finding shows that a high level of agreement with regard to assigned scores does not necessarily imply that assessors also agree with regard to underlying evidence and arguments. Only a small difference was found in variation in evidence and arguments between video episodes where assessors reached a high level of agreement in assigned scores and video episodes where they reached a low level of agreement. This finding shows that assessors can come to the same conclusion about teachers' coaching performance, based on different evidence and arguments. Furthermore, a low level of agreement with regard to assigned scores seems not only to be caused by a lack of agreement with regard to reported evidence and arguments; other processes may play a role here. For instance, it is possible that the process of assigning scores is not based exclusively on considerations relating to evidence and arguments reported on the score forms, but that in assigning scores other evidence and arguments, or emotions and personal beliefs, are also involved (Moss, 1994). Another conclusion is that more variation was found in arguments than in evidence. The reported arguments consisted mostly of inferences and judgments: statements at a higher level of abstraction. These inferences or judgments can be seen as interpretations of the observations that assessors made while collecting evidence. These results confirm those of Schutz and Moss (2004), who also found that assessors made very different, but legitimate interpretations based on the same evidence when judging portfolios. In making representations out of concrete evidence or observations, a system of constructs is involved. The (personal) associative connections in this system of constructs might explain the differences found in the (abstract) representations of the assessors (DeNisi, Cafferty, & Meglino, 1984; Carlston, 1992; 1994; Feldman, 1981; Landy & Farr, 1980). These results seem to indicate that even though assessors participated in an intensive training course of four training sessions, the training did not result in a completely shared understanding of constructs and associations related to competent coaching.

With regard to the second research question, it can be concluded that the assessors used a mix of concrete and abstract statements to justify the scores they assigned. This finding is in line with the results of a study by Van der Schaaf, Stokking, and Verloop (2005), who found similar results. In this study, assessors used mainly citations concerning concrete teacher behaviors as evidence, especially asking questions and providing feedback. The concrete questions and feedback were considered relevant evidence in the scoring guide and conceptual framework for competent coaching. Assessors seemed reasonably capable of identifying relevant, concrete evidence for competent coaching. In this part of the scoring process, only the slightest problems with regard to construct irrelevance and construct under-representation were encountered. The summary reports contained mainly inferences and judgments. The inferences mostly concerned teachers' behavior (18%) and the coaching situation (14%). These arguments were considered relevant arguments in the scoring guide and conceptual framework. The judgments in the summary reports concerned teachers' behavior (48% of all judgments) and also consequences for students (19% of all judgments). These arguments were also in line with the scoring guide and conceptual framework. The assessors focused more on teachers' behavior, and paid less attention to the consequences of teachers' behavior for the students, which was unexpected considering the performance levels that were formulated in terms of consequences for students. A plausible explanation for this finding is that teacher behavior is easier to observe and interpret for assessors than the consequences for students. During the training course, assessors indicated that they found it hard to judge the consequences for students. As noted earlier, the inferences and judgments reported in the summary reports were in line with the scoring guide and the conceptual framework, but related to different aspects of the conceptual framework: the coaching situation, teacher behavior, and consequences for students. In addition, also within these three aspects of competent coaching, assessors tended to focus on different sub-aspects. It appeared that instead of looking for evidence and arguments related to all of these aspects, assessors focus on only one or two. Furthermore, it is possible that the considerable variation in arguments that was reported earlier as a conclusion, can be attributed to assessors' focus on different aspects in the conceptual framework.

With regard to the third research question, it can be concluded that assessors did not report a lot of irrelevant evidence and arguments. Only 1% of the evidence and 4% of the arguments were irrelevant when compared with the conceptual framework. More construct-irrelevant citations were found when the assessors were urged to judge the coaching in a specific learning activity in the video episode. The results show that assessors not only reported evidence that referred to the coaching of this specific learning activity, but also referred to the coaching of other, construct-irrelevant, learning activities. Assessors reported slightly more irrelevant evidence during the scoring of the video episodes for which they reached the lowest level of agreement with regard to scores assigned to teachers' coaching performance. However, assessors reported slightly more irrelevant arguments during the scoring of the video episodes for which they reached the highest level of agreement with regard to scores assigned to teachers' coaching performance. A plausible explanation for these construct-irrelevant citations is that, in practice, the different kinds of learning activities are so interwoven and interrelated that it is hard for assessors to distinguish the evidence and arguments that relates to the coaching of a specific learning activity. It is possible that the distinction between the different learning activities can only be made in theory, and is less usable in practice. Another possible explanation is that the assessors need more training in distinguishing evidence and arguments related to the different learning activities.

What do these conclusions say about the reliability and validity of the designed assessment procedure? Can assessors' scoring processes be considered reliable and valid when so much variation in evidence and arguments were found? These questions seem to be related to another important question: Can the variation in evidence and arguments be explained by threats to reliability and validity, such as construct-irrelevant variance or construct under-representation (Messick, 1989), or do assessors report evidence and arguments that are consistent with the scoring guide and conceptual framework, but are just different, as was found in a study by Schutz and Moss (2004)? When construct-irrelevant variance and construct under-representation can be discovered in reported evidence and arguments, not only reliability, but also validity is at stake. It appears that the impact of construct-irrelevant variance on reported evidence and arguments was small; the reported evidence and arguments are mostly consistent with the scoring guide and related conceptual framework. The impact of construct under-representation was larger; assessors seemed to focus on only one or two aspects of the conceptual framework. These conclusions suggest that the variation in evidence and arguments was caused, at least to some degree, by construct under-representation. This may have had a negative influence on the validity and reliability of the scoring process, and thus on the validity and reliability of the performance assessment. Furthermore, the conclusions of this study suggest that more research is needed with regard to the assignment of scores to coaching performances in order to be able to get a complete indication of the validity and reliability of assessors' scoring process. The results show that a lack of agreement with regard to evidence and arguments did not automatically lead to a lack of agreement in assigned scores. This conclusion suggests that assigning scores is a process that is not entirely based on reported evidence and arguments. It is possible that assessors also took other evidence and arguments into account that they did not write down on the score forms. Such evidence and arguments could not be analyzed in this study. The aim of this study was to analyse evidence and arguments explicitly reported by the assessors. In order to get a realistic perception of the proportion of all construct-irrelevant variance and construct under-representation that plays a role in performance assessments, evidence and arguments that are not written down on score forms, but are also taken into account during the judgment process, should also be investigated. Furthermore, this study was focused on the kinds of evidence and arguments reported by assessors, and not on how assessors combined the different evidence and arguments in a judgment. Especially in the process of combining evidence and arguments, construct under-representation can occur. This part of the judging process will be a topic of our future research.

Another important question concerns the implications of the conclusions of this study for improving the reliability and validity of performance assessment procedures like video portfolios. First, in order to reduce the variation, especially in arguments, more attention should be given to creating a shared understanding of the conceptual framework (Frederiksen, Sipusic, & Sherin, 1998; Woehr & Huffcutt, 1994). During training, the discussion should be focussed more explicitly on relevant arguments that play a role in assigning scores. It is expected that a more shared system of relevant constructs can be built as a result of exchanging these arguments during discussions. Second, in order to reduce the threat of leaving out important aspects of the conceptual framework, assessors should be encouraged during training to concentrate on all aspects of the conceptual framework. Third, more attention should be paid during training to aspects of the conceptual framework that are not explicitly perceptible in the video portfolio, such as 'consequences for students'. Assessors indicated that they found it hard to make inferences and judgments about consequences for students. More discussions with regard to this topic during training may help assessors to get a grip on it, so that they become more inclined to make such inferences and judgments in scoring portfolios.

References

Aronson, E., Wilson, T.D., & Akert, R.M. (2007). *Social psychology* (5th ed.). Amsterdam: Pearson Education Benelux BV.

Baume, D., Yorke, M., & Coffey, M. (2004). What is happening when we assess, and how can we use our understanding of this to improve assessment? *Assessment & Evaluation in Higher Education, 29*(4).

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.

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

Carlston, D. (1992). Impression formation and the modular mind: The associated systems theory. In L.L. Martin & A. Tesser (Eds.), *The construction of social judgments.* Hillsdale, NJ: Erlbaum.

Carlston, D. (1994). Associated systems theory: A systematic approach to cognitive representations of persons. *Advances in Social Cognitions, 7*, 1-78.

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* (pp.453-494). Hillsdale, NJ: Erlbaum.

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

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.

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

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.

Dunbar, S.B., Koretz, D.M., & Hoover, H.D. (1991). Quality control in the development and use of performance assessments. *Applied Measurement in Education*, *4*, 289-303.

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.

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.

Kane, M.T. (2004). Certification testing as an illustration of argument-based validation. *Measurement, 2*(3). 135-170.

Kane, M.T. (2006). Validation. In R.L. Brennan (Ed.), *Education Measurement* (4th ed.). Westport: PraegerPublishers.

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. (1994). Performance assessment. Policy promises and technical measurement standards. *Educational Researcher, 23*(9), 4-14.

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

Messick, S. (1989). Validity. In R.L. Linn (Ed.), *Educational measurement* (3rd ed.). New York; MacMillan.

Messick, S. (1995). Validity of psychological assessment. Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist, 50*(9), 741-749.

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 inivironments 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.

Nijveldt, M. (2007). *Validity in Teacher Assessment: An Exploration of the judgments processes of assessors.* Doctoral dissertation. Leiden: ICLON Graduate School of Education.

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.

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.

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/.

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

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.

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

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

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.

Chapter 5

General conclusions and discussion

5.1 Overview of the study

The aim of the research presented in this dissertation was to contribute to the knowledge base pertaining to the reliability, generalizability, and validity of authentic performance assessment procedures in order to be able to improve the methodological quality of such procedures. As part of this dissertation an authentic performance assessment procedure was developed based on design principles that are expected to contribute to reliable, generalizable, and valid judgments. The assessment procedure was called 'video portfolios'. A video portfolio consists of a mix of sources of evidence that are expected to provide assessors with a complete picture of a teacher's competence. In this study, the video portfolios that were developed aimed at measuring the coaching competence of teachers who work in senior secondary vocational education. The main sources of evidence in a video portfolio are video episodes that represent a teacher's coaching performance (Frederiksen, Sipusic, Sherin, & Wolfe, 1998). For this, teachers were filmed on-the-job while they had coaching sessions with a group of students. The video episodes represent performance in an authentic context. In order to be able to score and judge teachers' coaching performance in the video episodes in a valid way, also other sources of evidence were included in the video portfolios. These sources concerned information about the learning task the students worked on during a video episode, information about students' progress with regard to completing the task, the students' backgrounds, the teachers' background, interviews with the teachers about the decisions that underlied their actions, and interviews with students about the perceived impact of the teachers' behavior on their work. In addition to these sources of evidence, information was added to the video portfolios about educational materials that were used and students' products that were discussed during the video episodes. The central research question of this dissertation was: to what extent are judgments based on video portfolios reliable, generalizable, and valid? In order to answer this research question, three studies were conducted that focused on different aspects of this research question.

Study 1 is a small-scale study, which reports on the design, development and use of video portfolios. In this study, two important aspects of reliability and validity were investigated: interrater agreement and aspects in the design of the video portfolios that stimulate of hinder assessors in making valid interpretations and judgments.

First, an assessment procedure called 'video portfolio' was developed. The construction of the video portfolios in this study started with conducting a detailed domain analysis concerning teachers' coaching competence in senior secondary vocational education. Based on this analysis, a solid scoring guide and conceptual framework were elaborated which were expected to assist assessors in making valid interpretations and judgments with regard to teachers' coaching competence. With the aid of a professional production team, video portfolios were constructed. Various sources of evidence were collected about a series of four coaching sessions spread over the four weeks that students worked on one complex task. In the construction of the video portfolios, serious efforts were made to ensure issues of content representation in terms of relevant coaching situations and in terms of the task processes on the part of the teacher. Furthermore, also a scoring procedure was developed which was expected to assist assessors in making reliable and valid interpretations and judgments. Assessors were asked to judge the video episodes based on a detailed scoring procedure starting with scoring specific aspects of the performance according to criteria and performance levels for competent coaching. Subsequently, assessors were asked to assign a score to the whole performance shown in a video episode and to the coaching performance across video episodes (overall score). Finally, an assessor training was developed in which assessors were trained in using the scoring guide, conceptual framework, and scoring procedure. After the development of the video portfolios, trained assessors were asked to score the video portfolios according to the scoring procedure.

Second, in order to get an indication of the interrater agreement, assigned scores to video episodes and assigned overall scores were collected and the interrater agreement was determined. Furthermore, a semi-structured interview was carried out with all assessors in order to obtain information concerning aspects of the assessment procedure that stimulated or hindered assessors in making valid interpretations and judgments. The main findings of this study were that an acceptable to high level of agreement between assessors was found which indicates that assessors arrived at corresponding scores. However, assessors indicated that mastering the scoring procedure takes time and energy. They perceived the assessor training as a necessary condition for applying the scoring procedure in the right way. Several factors were found to be helpful for assessors in making valid interpretations and judgments. Assessors indicated that the following factors assisted them in making valid interpretations and judgments:

- a scoring guide with descriptions of learning activities and related coaching interventions, because these tools direct assessors towards relevant aspects of the coaching performance;
- summaries of what happened during the coaching situation, because these summaries direct assessors also towards relevant aspects of the coaching performance;
- context information, especially the interview with the teacher and the student(s), because it helped understanding teachers' behavior and the consequences for students;
- straightforward coaching situations, i.e., situations referred to by the asssessors as 'clear' and 'less complex'; characteristics of those situations are, for example: a clear match between a teacher's intentions and behavior, coaching in a specific learning activity that clearly differs from the coaching in other learning activities, and the need of support by students in only one specific learning activity.

Some disabling factors were also indicated by assessors, namely:

- a single video episode appeared to be difficult to score, because a single video episode only shows a fragment of what happens between teacher and student(s);
- video episodes that are longer than 15 minutes did not seem to contribute to valid interpretations and judgments, because it was hard for assessors to concentrate longer than 15 minutes and, according to assessors, no crucial evidence revealed after 15 minutes;
- it appeared sometimes to be difficult to distinguish coaching on performance level 2 from coaching on performance level 3, this was the critical distinction between a negative and a positive score;
- the degree to which teachers' coaching was practice-oriented coaching could not be judged in a valid way, because teachers barely or not showed any behavior with regard to this criterion; consequently, in judging teachers' coaching with regard to this criterion, assessors could only rely on negative evidence in terms of missed opportunities.

In study 2, the reliability of assessors' scoring and the generalizability of judgments were investigated based on several quantitative analyses concerning scores assigned to video episodes and overall scores. The analyses with regard to assessors' scoring included the examination of tendencies in assessors' assigned scores, interrater agreement, and the generalizability of scores across assessors. The analyses with regard to the generalizability of judgments were based on a ranking of the video episodes: the video episodes that elicited the most similar scores were placed high in the ranking order and the video episodes that elicited the most varying scores were placed low in the ranking order. Especially the video episodes that elicit the most varying scores are a threat to the generalizability across video episodes. Furthermore, for each video episode it was determined to what extent the score assigned to the specific video episode matched the scores assigned to the other video episodes. The main findings of this study were that assessors' scoring seemed to be supported by the design of the assessment procedure. In general, the assessment procedure enabled reliable scoring by assessors. The results show an acceptable level of agreement for the video episodes and a high level of agreement for the assigned overall scores; thus reliability could be realized for the assigned scores. The generalizibility of scores across assessors was also high. The results indicate that when two assessors participate in the assessment procedure, an acceptable level of interrater agreement can be realized. However, scoring tendencies appeared to influence assessors' scoring; assessors did not judge the different teachers equally leniently or severely. Furthermore, assessors who knew the colleagues to be judged, were inclined to assign extreme lenient or severe scores. The main findings with regard to the generalizability of scores across video episodes are that the selection of key situations as videos episodes may have had a positive effect on the generalizability of scores to the intended universe of video episodes. The agreement between scores assigned by assessors to the same 'type' of video episodes (video episodes where teachers' coached on cognitive, meta-cognitive, affective, or collaborative learning activities) was predominantly acceptable to high, especially for the video episodes of teacher 1 and 2 and video episodes concerning coaching in cognitive learning activities. Only the agreement between scores assigned by the assessors to the video episodes concerning coaching in affective learning activities appeared to be problematic.

In study 3, the validity of assessors' scoring process was investigated. A qualitative content analysis was conducted on evidence and arguments that assessors reported on score forms to justify their assigned scores. Based on this analysis, the impact of construct-irrelevant variance and construct under-representation of assessors' judgments was examined. A considerable amount of variation was found in the reported evidence and arguments. Furthermore, more variation was found in arguments than in evidence. Assessors used a mix of concrete and abstract statements; concrete statements were predominantly used as evidence and abstract statements predominantly as arguments. The evidence and arguments were consistent with the conceptual framework, so that little construct-irrelevant evidence and arguments were reported by the assessors. However, the assessors scoring seemed to be influenced by construct under-representation, because of their tendency to focus on only one or two aspects of the conceptual framework when interpreting and judging video episodes, instead of all aspects.

5.2 Conclusions and discussion

In this section, the main conclusions are presented and discussed. The central research question of this dissertation was: to what extent are judgments based on video portfolios reliable, generalizable, and valid? In section 5.2.1 the conclusions with regard to the reliability of the assessment based on video portfolios are presented and discussed, in section 5.2.2 with regard to the generalizability of the assessment based on video portfolios, and in section 5.2.3 with regard to the validity of the assessment based on video portfolios.

5.2.1 Reliability of judgments based on a video portfolio

The reliability of scores assigned to (aspects of) video portfolios was examined in study 1 (based on six assessors) and in study 2 (based on 12 assessors). The agreement among assessors with regard to evidence and arguments was examined in study 3 (based on 12 assessors). From these studies, five main conclusions can be drawn with regard to the reliability of the authentic performance assessment based on video portfolios.

Conclusion 1: Assessors reached an acceptable to high level of agreement with regard to the assigned (overall) scores based on video portfolios.

Although it is often claimed that it is difficult to realize agreement among raters in authentic performance assessments (Baume, & York, 2002; Delandshere, & Petrosky, 1998; Gipps, 1994; Moss, 1994), the results of the studies in this dissertation show that it is possible to reach an acceptable to high level of interrater agreement based on video portfolios. It can be assumed that the design principles used in the construction of the video portfolios supported the assessors' scoring and, thus, contributed to the interrater agreement. The results from the interview with the assessors from study 1 sustained this assumption. Assessors perceived especially the scoring guide with descriptions of learning activities and concrete examples of coaching interventions as helpful in making judgments. They indicated that these descriptions and examples directed their attention to the relevant aspects of the performance. Also the detailed description of the performance levels were perceived as helpful, only the distinction between performance level 2 and 3 was sometimes hard to make. The distinction between performance level 2 and 3 is the critical distinction between a negative and a positive judgment in the designed assessment procedure. Apparently, assessors found it especially difficult to make decisions that are around these performance levels. This finding suggests that in case of making high-stakes decisions some adjustments need to be made in the assessment procedure. During the assessor training specific attention should be given to aspects of coaching that are typical for coaching on score level 2 and typical for coaching on score level 3, so that assessors will be better able to make a decision with regard to coaching on score level 2 and score level 3.

Conclusion 2: Assessors reached a higher level of agreement for the overall scores than for the scores they assigned to single video episodes.

A higher level of interrater agreement was found for overall scores when compared to scores assigned to separate video episodes. Whereas assessors sometimes varied in their judgments concerning performance in single video episodes, they agreed on teachers' level of performance across different video episodes. The interview results from study 1 provide more information with regard to this phenomenon. In the interview, assessors indicated that it is harder to interpret and judge single video episodes, because it shows only a fragment of what happens between teacher and student(s). Although several sources with context information were added to the video episodes in order to provide assessors with a complete picture of the teachers' performance (Heller, Sheingold, & Myfords, 1998; Schutz, & Moss, 2004), the single video episodes can sometimes have a too fragmented character. Assessors indicated that based on five to six video episodes they could get a pretty clear view of teachers' performance.

Conclusion 3: Two assessors are needed to establish an acceptable level of interrater agreement.

Study 2 shows that an acceptable level of interrater agreement can be established by using only two to three assessors. Although the use of more assessors in an assessment procedure contributes to a higher interrater agreement, the agreement based on two assessors is acceptable and does not improve much by adding more assessors. This is in line with results found by Dunbar, Kortez, and Hoover (1991). Important to note is that only an acceptable level of interrater agreement based on two to three assessors can be established under the same conditions as in this study. In this study, several measures were taken in the design of the assessment procedure to ensure reliable scoring, such as the use of a scoring guide and a conceptual framework, a detailed scoring procedure, and an assessor training.

Conclusion 4: Assessors' scoring showed scoring tendencies.

Based on study 2, a specific threat to reliable scoring was detected. It appeared that scoring tendencies occured in the process of assigning scores. The results of the study show that assessors did not judge the different teachers in an equally lenient or severe way. This finding shows that, at least to some extent, assessors' scoring was inconsistent for which no unequivocal explanation can be given. It might be that it was hard to judge the teachers in a consistent way, because they were filmed in different contexts. It could also be that assessors were influenced by personal biases or preferences of a specific coaching style (Gipps, 1994; Moss, 1994). Study 2 showed that especially colleagues of the teachers who were filmed and included in the portfolios were suffering from inconsistent scoring. These assessors scored their colleagues either extreme leniently or extreme severely. From the literature it is known that assessors who are close to the person judged, will be tempted to judge leniently (Aronson, Wilson, & Akert, 2007). This tendency would explain why some assessors who judged their colleagues assigned extreme lenient scores. However, the results show that some assessors who judged their colleagues also assigned more severe judgments. This cannot be explained by known scoring tendencies from literature and might have to do with personal characteristics of the assessor(s).

Conclusion 5: Assessors based their scores on mutually differing evidence and arguments. More variation was found between assessors with regard to arguments than with regard to evidence.

Although assessors reached an acceptable to high level of agreement with regard to assigned (overall) scores, assessors did not base their scores on similar evidence and arguments. This finding shows that a lack of agreement in reported evidence and arguments does not automatically lead to a lack of agreement in assigned scores. There can be three explanations for the variation in evidence and arguments found among assessors. A first explanation might be that assessors just differed in the way they arrived at the (same) assigned score. A second explanation comes from the results of study 3. This study shows that assessors appeared to focus on different aspects of the conceptual framework when they interpreted and judged a video episode. Some assessors used evidence and arguments that were related to teachers' behavior and the context of the coaching situation, while other assessors focused more on evidence and arguments that were related to consequences for students. This tendency to focus on different types of evidence and arguments explains the variation found in evidence and arguments. A third explanation can be that the process of assigning scores is not solely based on reported evidence and arguments on score forms. It might have been that assessors based their assigned score not only on evidence and arguments that they reported, but also on evidence and arguments that they had in mind, but did not write down on the score forms. Personal beliefs and emotions may also have had an impact on the process of assigning scores (Gipps, 2004; Moss, 2004).

From the results of study 3 it appears that variation between assessors especially arises in formulating (abstract) arguments. This result is in line with results found by Schutz and Moss (2004), who concluded that assessors can make very different, but legitimate interpretations based on the same evidence when judging portfolios. This finding might be explained by the fact that especially in interpreting observations a system of constructs is involved in which (personal) associative connections exist (Carlston, 1992, 1994; DeNisi, Cafferty, & Meglino, 1984; Feldman, 1981; Landy & Farr, 1980) and which influence the assessors' evaluation of observations. Considering the variation that was found in arguments, the assessors seemed not to evaluate teachers' coaching performance based on a totally shared understanding of constructs and associations concerning competent coaching. It might be that more training sessions are needed to establish a shared system of constructs, than the four sessions that were used in this study.

5.2.2 Generalizability of judgments based on a video portfolio

The generalizability of scores to the intended universe of video episodes was examined in study 2. Important to note is that based on this study, it was only possible to describe tendencies with regard to the generalizability. No hard conclusions could be drawn with regard to the minimum number of video episodes needed to reach an acceptable level of generalizability. Furthermore, the interview study from study 1 provided more information about interpreting and scoring different video episodes. Two main conclusions concerning the generalization of judgments can be drawn.

Conclusion 6: The scores assigned to video episodes of some teachers were better generalizable than scores assigned to video episodes of other teachers.

The results of study 2 show that the generalizability of scores assigned to video episodes of teacher 1 and teacher 2 were better generalizable to a universe of video episodes than scores assigned to video episodes of teacher 3 and 4. The generalizability of scores assigned to video episodes of teacher 3 was the lowest. Based on the results of study 2, it is hard to predict the reason why the scores assigned to video episodes of teacher 1 and 2 could be better generalized than the scores assigned to video episodes of teacher 3. It might be that the teachers 1 and 2 reacted more consistent to the coaching situations than teacher 3. But it might also be that the assessors scored teacher 1 and 2 in a more consistent way than teacher 3. In study 2, the lowest level of interrater agreement was found for teacher 3. This result shows that also in study 2, problems with the scoring of the video portfolio of teacher 3 were detected. Furthermore, assessors reported in study 1 that especially video episodes that were longer than 15 minutes were hard to score. The video episodes that were included in the video portfolio of teacher 3 were predominantly longer than 15 minutes, which might have influenced the assessors' scoring. In study 1, assessors also reported that complex video episodes were hard to score. They indicated that especially video episodes in which students needed support in multiple learning activities at once and where the teacher was coaching on several learning activities at the same time were hard to score. The students in the video episodes of teacher 3 had severe motivation problems (which pertain to affective learning activities), which also led to problems with regard to collaborative processes (which pertains to learning activities with regard to collaborative learning). These two problems were present in all video episodes of teacher 3 and the teacher was expected to address these hard problems. It seemed that the video episodes of teacher 3 are typical examples of what the assessors indicated as 'complex video episodes', a factor thus that influenced the scoring of the video portfolio of teacher 3. These findings suggest that in order to be able to generalize scores to a universe, it is recommended to include video episodes in the video portfolios that are less complex and last no longer than 15 minutes. Sometimes complex video episodes cannot be avoided. In that case, more video episodes could be included in a video portfolio to realize generalizability or more assessors could be used to judge the complex video episodes.

Conclusion 7: The scores assigned to video episodes concerning coaching of some learning activities were better generalizable than scores assigned to video episodes concerning coaching of other learning activities.

The results of study 2 show that the generalizability of scores assigned to video episodes concerning coaching in cognitive learning activities could be generalized to the universe of video episodes, but the generalization of scores assigned to video episodes concerning coaching in affective learning activities appeared to be problematic. Based on results of study 2, it is hard to predict why the scores assigned to the video episodes concerning cognitive learning activities can be better generalized to other video episodes than scores assigned to video episodes concerning affective learning activities. It may be that the coaching in affective learning activities happens very subtle and is interwoven with coaching in other learning activities. This might make it very hard for assessors to score the coaching in affective learning activities in a consistent way, which is in line with the results of study 1 where assessors reported that especially video episodes where the teacher coached on multiple learning activities were hard to score. This finding suggests that, before using the designed assessment procedure in practice, some adjustments have to be made to improve the generalizability of scores assigned to video episodes in which the teacher coaches on affective learning activities. It is expected that especially the inclusion of more video episodes with regard to the coaching of these learning activities will improve the generalizability.

5.2.3 Validity of judgments based on a video portfolio

The validity of scores assigned to (aspects of) video portfolios was examined in study 1 and 3. In study 1, assessors were interviewed about factors that stimulated or hindered them in making valid interpretations and judgments. In study 3, a thorough investigation of construct-irrelevant variation and construct under-representation in reported evidence and arguments was carried out. Based on these studies three main conclusions can be drawn.

Conclusion 8: Assessors perceived the context information that was included in the video portfolios as indispensable background information for validly judging video episodes.

In the interview in study 1, assessors reported that particularly the interviews with the teachers and the students were perceived as indispensable background information for making valid interpretations and judgments. These information sources informed assessors about teachers' decisions that underlied their performance and about the impact of teachers' behavior on students. The importance of knowledge about teachers' underlying decisions is supported by a study of Schutz and Moss (2004) in which they focused on underlying intentions. It appeared that when assessors were not informed about teachers' intentions, assessors make assumptions for themselves about their intentions in order to be able to interpret and judge teachers' performance.

Conclusion 9: Assessors were able to use evidence and arguments in scoring the video portfolios that were consistent with the conceptual framework.

Although a lot of variation was found between assessors with regard to evidence and arguments (conclusion 5), the variation seemed not to be caused by the use of irrelevant evidence and arguments. Most of the scoring process was based on construct relevant evidence and arguments. This was also found in other studies (Heller, Sheingold, & Myford, 1998; Nijveldt, 2007). Only 1% of the evidence and 4% of the arguments were irrelevant compared to the conceptual framework. In addition, little more construct-irrelevant evidence and arguments were found in assessors' judging of the coaching on a specific learning activity. It seemed that assessors had trouble with judging the coaching on a specific learning activity and, when doing that, to exclude judging of the coaching on other learning activities. A plausible explanation for the use of evidence and arguments that are related to the coaching on other learning activities is that, in practice, the coaching of different types of learning activities are so interwoven and interconnected that it is hard for assessors to judge only the coaching on a single learning activity. This explanation implies that a strict distinction between several types of learning activities is for assessors less useful in practice. However, it could also be that assessors just needed more training in judging the coaching on a specific learning activity in order to be able to identify evidence and use arguments that are related to the coaching of the learning activity that assessors were expected to judge.

Conclusion 10: The validity of assessors' scoring may have been negatively influenced by assessors' focusing on only one or two aspects of the framework instead of all aspects.

Study 3 reveals that although assessors reported evidence and arguments that were consistent with the conceptual framework (conclusion 9), assessors tended to focus on different aspects of the conceptual framework. The evidence and arguments that were reported by assessors were related to the coaching context, teachers' behavior, or consequences of teachers' behavior for students. It appeared that instead of looking for evidence and arguments related to all these three perspectives, assessors reported only evidence and arguments that were related to one or two aspects. This finding suggests that assessors left out some perspectives on competent coaching in the scoring process. This points to under-representation of aspects in the framework. The exclusion of some aspects threatens the validity of the scoring process. Assessors should be instructed and trained more explicitly to include all the aspects of the framework in assigning scores to teachers' coaching performance. This finding is related to conclusion 5; the assessors' focus on different perspectives of competent coaching may explain the large variation that was found in reported evidence and arguments on score forms.

5.3 Limitations of the study

In this section, three aspects of the studies are discussed that limit the conclusions: (a) the number of video episodes that were included in the study, (b) the focus on reported evidence and arguments on score forms, and (c) combining evidence and arguments to a judgment.

Size of the sample of video episodes

Due to the small sample of video episodes used in this study, no proper generalizability study (Brennan, 2001) could be conducted to determine the exact number of video episodes needed to reach an acceptable level of generalizability. The small sample of video episodes was chosen, because the construction of the video portfolios according to design principles in the literature was complex and time consuming. In order to construct a solid performance assessment, the video portfolios were constructed very precise. Furthermore, these video portfolios were new, therefore, we started to create and test these portfolios on a relatively small scale. Alternatively, two analyses were conducted in order to obtain information about the generalizability across video episodes. In the first analysis, video episodes that were scored differently by different assessors were identified. These video episodes have a negative effect on generalizing across video episodes. In the second analysis, it was determined to what extent a score assigned to a specific video episode matched the scores assigned to other video episodes. Video episodes with matching scores have a positive effect on generalizing across video episodes.

Focus on reported evidence and arguments on score forms

The validity of the scoring process of assessors was investigated in detail in study 3. In that study, evidence and arguments that assessors used to justify an assigned score were examined for construct-irrelevant variance and construct under-representation. The analyses were conducted on evidence and arguments that assessors reported on the score forms. However, by relying on only reported evidence and arguments entails the danger that not all evidence and arguments that play a role in assigning scores are analyzed. The impact of construct-irrelevant variance and construct under-representation on the validity of assessors' scoring might have been larger than was found in study 3. After all, the construct-irrelevant variance and construct under-representation for evidence and arguments that assessors used, but not wrote down on the score forms, were not covered in this study.

Combining evidence and arguments to a judgment

In study 3, evidence and arguments that were reported on score forms were analyzed for construct-irrelevant variance and construct under-representation. These analyses focused on what evidence and arguments assessors reported on score forms. However, construct under-representation can also have an impact on the process of weighing and combining evidence and arguments by placing an inappropriate emphasis on specific evidence and arguments. This part of the scoring process is not investigated in our study. By leaving out this aspect of the scoring process, it could be that the magnitude of construct under-representation may in fact have been larger than was found in study 3.

5.4 Suggestions for future research

Three directions for future research are proposed: (a) research that focuses on the extrapolation to performance outside the assessment context, (b) research that focuses on teachers' learning based on the assessment procedure, and (c) research that focuses on characteristics of assessment tasks.

Extrapolation to performance outside the assessment context

The aim of the studies presented in this dissertation was to investigate the internal validity of the performance assessment (Lissitz & Samuelson, 2007); the focus was on assessors' scoring and the generalization of scores to a universe of scores. However, another vital aspect of validity is the relation between assessment scores and external measures (extrapolation inference; Kane, 2006). In the design of the video portfolios, several measures were taken to warrant the extrapolation to performance outside the assessment context: (1) high-fidelity assessment tasks were used that represent the complex situations that teachers face in practice, (2) domain coverage was expected to be realized by including ten video episodes in each video portfolio that covered the coaching in different learning activities, and (3) the video portfolios encompassed four weeks in which students worked on one complex task. However, the contribution of these design principles to the possibility of extrapolation of scores was not investigated in this study and might be the topic of future research. This type of research includes a job analysis that shows what situations teachers face in practice and how often. Subsequently, the sample of assessment tasks should be tuned to this job analysis in order to realize content coverage. Based on this type of research, the design principles to ensure extrapolation can be adjusted and refined in order to further improve the (methodological) quality of performance assessments.

Research that focuses on teachers' learning based on the assessment procedure

The studies presented in this dissertation were conducted in order to investigate to what extent teachers' coaching competence can be determined in a valid way based on the assessment procedure constructed. It would also be interesting to investigate to what extent the constructed assessment procedure contributes to teachers' learning with regard to coaching students (i.e., Darling-Hammond & Snyder, 2000; Lusttick & Sykes, 2006). In other words, can the portfolios be used for formative assessment purposes? Especially the summaries in which assessors report evidence and arguments that explain why they assigned the specific score to teachers' coaching competence can be very helpful in teachers' development towards an expert coach. Furthermore, the teachers who acted as assessors, felt that they had learned a lot about coaching during the assessor training. Assessors indicated that especially discussing the coaching performance of the teachers in the video episodes, helped them to reflect on their own coaching in practice.

Characteristics of the video episodes

The studies presented in this dissertation showed that the generalizability of scores assigned to some video episodes is better than for other episodes. Furthermore, assessors indicated that some video episodes are easier to score than others. These findings raise questions like: 'what makes scores assigned to some video episodes better generalizable than others?' and 'what makes some video episodes easier to score than others?' Based on the results of the studies in this dissertation, some indications are obtained with regard to these topics. It appeared that video episodes that are 'less complex' are easier to score and generalize. Further research is needed in order to answer these questions in more detail. Furthermore, not only research that focuses on assessors' scoring is needed, but also research in which the characteristics of assessment tasks are systematically compared (in relation to assessors' scoring).

Insights obtained by such research can be used to formulate additional design principles for the construction of assessment tasks in performance assessment procedures.

5.5 Implications for assessment practices

A number of practical implications can be derived from the studies described in this dissertation, which can be used to warrant and improve the reliability, validity, and generalizability of authentic performance assessments such as video portfolios.

Design of the assessment procedure

Video portfolios as a method for accomplishing a reliable, valid, and generalizable performance assessment seems to be promising. The studies in this dissertation show that the design principles that were proposed in literature and used for the construction of the assessment procedure generally went together with positive results concerning assessors' scoring and generalizability. Therefore, it is recommended to use the following design principles when developing an assessment procedure for assessing teachers' competence:

- a scoring guide that includes criteria, performance levels and concrete examples of competent and incompetent performance (Fredriksen, Sipusic, Sherin, & Wolfe, 1998);
- a combination of a literature-based as well as practice-based scoring guide (Uhlenbeck, 2002);
- a scoring guide that contains only aspects that distinguish competent from (in)competent performance (Dwyer, 1993; Kagan, 1990);
- criteria and performance levels formulated in terms of what a teacher should achieve in terms of the consequences of teachers' behavior for students (Darling-Hammond & Snyder, 2000; Dwyer, 1998; Gipps, 1994; Haertel, 1991; Uhlenbeck, 2002);
- multiple information sources in order to cover all aspects of teaching (Beijaard & Verloop, 1996; Dwyer, 1998; Uhlenbeck, 2002);
- a detailed scoring procedure starting with the scoring of specific aspects of the performance and, next, building a judgment of the whole performance (Klein & Stecher, 1998);

- an assessor training consisting of several sessions in which attention is paid to creating common conceptualizations concerning competent performance and to categorizing performance into the same performance levels (Woerh & Huttcuff, 1994);
- standardizing assessment tasks to some extent (Kane, 2006).

Results from study 1 show that especially the descriptions of learning activities and concrete examples of coaching interventions in the scoring guide were perceived as very helpful in scoring the coaching performance. It helped the assessors to direct their observations to the relevant aspects of the performance. The inclusion of context information in the video portfolio also contributed to a better understanding and thus scoring of the performance shown in the video episode. Especially the interviews with the teacher and student(s) were perceived as indispensable background information. However, also some factors were found to have a negative effect on assessors' scoring such as video episodes that were longer than 15 minutes and video episodes that concerned complex coaching situations. Assessors referred to complex coaching situations as situations in which the teacher coached on several learning activities at the same time or situations in which students had problems with multiple learning activities. The studies showed that such video episodes may be a threat to valid and reliable scoring and to the generalizability across video episodes. It is therefore recommended to include video episodes in video portfolios that are less complex and which last no longer than 15 minutes. In case where complex episodes cannot be avoided, it is suggested to let these episodes be judged by a larger number of assessors or to provide assessors with more video episodes of that specific teacher.

Assessors

An important implication of the studies is that the use of two assessors in the assessment procedure should be enough to realize an acceptable level of reliability given that they have had a detailed assessor training and that the conditions in the assessment procedure are similar to those in this study. This is important, because in practice it is not possible to use twelve assessors like in the design that was used in this dissertation. Furthermore, it appeared that colleagues of the teacher to be assessed should better not be used as assessors, because they are inclined to make extreme judgments.

Assessor training

The assessor training used in this assessment procedure was based on elements of the Frame-Of-Reference training and on elements of the Rater-Error-Training (Woerh & Huttcuff, 1994). In addition to the content of the assessor training as recommended in literature, the results in this dissertation have also some implications for improving such trainings. First, it appeared that assessors found it hard to distinguish performance on level 2 from level 3. The distinction between level 2 and 3 was the critical distinction between a negative and a positive judgment in our assessment procedure. This finding suggests that during the assessor training more attention should be given to characteristics of coaching on level 2 and coaching on level 3 to be better capable of making a fair judgment. Second, it appeared that assessors were inclined to use only one or two aspects of the conceptual framework in scoring teachers' coaching performance. In order to overcome this phenomenon, assessors should be encouraged to use all aspects at the same time. Explicit feedback concerning the use of the conceptual framework in this way during the training might be an effective measure. Third, in order to reduce the considerable variety in especially arguments that was found in study 3, more attention should be given to the realization of a shared understanding with respect to the conceptual framework. Discussions during the training should be more explicitly focused on relevant arguments that play a role in assigning scores. By exchanging these arguments among assessors, it is expected that a more shared system of constructs will be build. Fourth, it appeared that assessors were more inclined to use evidence and arguments that pertained to teachers' behavior than to consequences of teachers' behavior for students. This was a rather surprising finding, because the performance levels were formulated in terms of consequences for students. A plausible explanation for this finding is that it is easier for assessors to evaluate teachers' behavior, because this is better perceptible than consequences for students. In order to stimulate assessors to make interpretations and judgments concerning consequences for students, discussions with regard to this topic can take place during the training so that assessors are explicitly trained in making these types of interpretations and judgments.

Final Remark: practical feasibility of video portfolios

The video portfolios designed in this study, were primarily constructed in order to investigate proposed design principles in the literature and in order to obtain new insights in processes and factors that affects the reliability, generalizability, and validity of performance assessments such as video portfolios. The practical feasibility of the video portfolios had less priority in the design of the video portfolios. The idea behind this approach was to investigate the reliability, generalizability, and validity under 'ideal conditions'. The assumption was that when the methodological quality of the assessment could not be ensured under ideal conditions, that it will be impossible to realize this in practice.

The video portfolios as designed in this study were not primarily designed for direct application in practice, but first of all for research purposes. However, as indicated in the previous section, some aspects of the assessment procedure can be directly used in practice. The scoring guide and conceptual framework pertaining to competent coaching is an example of such an aspect and also the scoring procedure and assessor training developed in this study can be used in practice. However, it is recommended to think about what teacher competences should be assessed based on video portfolios and what competences not. Assessors reported that the scoring procedure was time consuming especially in the beginning, so it is advisable not to use video portfolios for assessing all teacher competences in practice, but only a limited number of important ones. In order to use video portfolios in practice, some aspects need further investigation with regard to the practical feasibility. This concerns especially the recording of the videos in collaboration with a professional company and the organization of evidence in a multimedia environment. For these aspects of the assessment procedure, it should be investigated in what way costs and time can be reduced.

References

Aronson, E., Wilson, T.D., & Akert, R.M. (2007). Social psychology (5th ed.). Amsterdam: Pearson Education Benelux BV.

Baume, D., & Yorke, M. (2002). The reliability of assessment on a course to develop and accredit teachers in higher education. *Studies in Higher Education, 27*(1), 7-25.

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

Brennan, R.L. (2001). *Statistics for social sciences and public policy*. New York: Springer.

Carlston, D. (1992). Impression formation and the modular mind: The associated systems theory. In L.L. Martin & A. Tesser (Eds.), *The construction of social judgments*. Hillsdale, NJ: Erlbaum.

Carlston, D. (1994). Associated systems theory: A systematic approach to cognitive representations of persons. *Advances in Social Cognitions, 7*, 1-78.

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

Delandshere, G., & Petrosky, A.R. (1998). Assessment of complex performances: Limitations of key measurement assumptions. *Educational Researcher, 17*(2), 14-24.

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.

Dunbar, S.B., Koretz, D.M., & Hoover, H.D. (1991). Quality control in the development and use of performance assessments. *Applied Measurement in Education*, *4*, 289-303.

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. (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.

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.

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

Kane, M.T. (2006). Validation. In R.L. Brennen (Ed.), *Education Measurement* (4th ed.). Westport: PraegerPublishers.

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.

Lissitz, R.W., & Samuelson, K. (2007). A suggested change in terminology and emphasis regarding validity and education. *Educational Researcher*, *36*(8), 437-448.

Lusttick, D., & Sykes, G. (2006). National board certification as professional development: What are teachers learning? *Education Policy Analysis Archives, 14*(5). Retrieved August, 10, 2006 from http:/epaa.asu.edu/epaa/v14n5.

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

Nijveldt, M. (2007). *Validity in Teacher Assessment: An exploration of the judgments processes of assessors*. Doctoral dissertation. Leiden: ICLON Graduate School of Education.

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/.

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.

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

Apendices

Appendix 1 Aspects of learning activities

	Aspects of learning activities	Coaching interventions
Cognitive learning activities	**Orientation towards the complex task** The orientation towards the complex task involves determining what goal should be achieved and what domain-specific task requirements and task conditions are involved. Students need to take this domain-specific context into account while engaging in realistic planning and choosing an appropriate approach to fulfilling the task. It all comes down to learning to consider different approaches to realizing a product that complies with the requirements. Students have to learn to choose the best approach in the specific context. The orientation towards the task also includes determining what is needed to fulfill the task in terms of domain-specific knowledge, skills, materials, etc.	Questions - What are important task requirements that you should take into account? - Why did you choose this approach? - What are the advantages of this approach in this situation? Feedback - Alerting students to important task requirements that they should take into account. - Alerting students to knowledge that they need to fulfill the complex task. - Giving examples of appropriate approaches in different contexts. - Proposing alternative approaches.
	Searching for and organizing relevant information Searching for and organizing relevant information involves determining what information is needed to complete the complex task. Students need to know where relevant information can be found, and they should be able to judge the quality of the information obtained. Students should also be capable of selecting and organizing information.	Questions - How did you determine what information you needed to complete the complex task? - Where did you search for relevant information? - Are there any other sources where you can find information about x? - Is this information relevant in this case? - Do you think this information is up to date? Feedback - Giving clues as to where to find relevant information. - Giving examples of how to find relevant information.

Appendix 1 Aspects of learning activities (Continued)

	Aspects of learning activities	Coaching interventions
Cognitive learning activity	**Comprehending and using relevant subject matter** This learning activity involves comprehending and using technical facts, constructs, formulas, rules, routines, norms, and techniques. Students working on complex tasks run into problems like misconceptions of constructs or principles, or discover they are not acquainted with relevant facts, rules, or norms that should be applied. Students should be supported in this.	Questions - What is a cavity slat? - What kind of materials do you need to build a dam? - What calculation do you use to determine the volume of a ditch? - What is the standard height of a door? - If you choose to place a concrete floor, what are the consequences for the foundation? Feedback - Referring students to relevant theory.
Meta-cognitive learning activities	**Planning** Students must be able to make a complete plan. They should allocate a realistic amount of time to different subtasks and make a realistic estimation of necessities (in terms of information and materials). A plan should also contain information on the approach chosen to accomplish the task.	Questions - What is the advantage of using a plan? - In what way can you estimate the time you need for x? - Where in the plan can I see how much time I can spend on building y? Feedback - Alerting students to possible bottlenecks in the planning. - Alerting students to omissions in the plan. - Giving examples of how to estimate how much time is needed for a specific job.
	Evaluating This learning activity involves being able to indicate what went well and what went wrong while carrying out the complex task. Among other things, students should pay attention to the extent to which they have reached personal goals: what have I learned and how will I do this next time?	Questions - What went well and what did not go so well in completing the task? - What things are you going to do different in the next task? How are you going to do that next time? - What did you learn while you worked in this complex task? Feedback - Alerting individual students to specific things they should pay attention to in the following task. - Alerting individual students to their progress in x.

Appendix 1 Aspects of learning activities (Continued)

	Aspects of learning activities	Coaching interventions
Meta-cognitive learning activity	**Monitoring and adjusting** Students should check whether the proceedings are in line with the plan made, and whether the chosen approach will lead to the desired outcomes. Based on the outcomes of the check, students have to decide whether adjustments to the approach are needed.	Questions - Which (sub)tasks are finished and which (sub)tasks are not yet finished? - How can I see in the plan which (sub)tasks should be finished? - You haven't finished task x yet; what are the consequences for meeting the rest of the plan? - To what extent does the product meet the requirements? - How can you make sure to finish the (sub)task in time? - If you don't adjust your approach, what are the consequences? - How can you make sure that you meet the product requirements? Feedback - Alerting students to the importance of checking the plan with regard to finishing the (sub)tasks in time. - Giving examples of how to check whether you meet the planning or not. - Alerting students to the requirements that the product should meet. - Giving examples of how to adjust the planning so that the task can be completed in time. - Giving clues as to how to adjust the approach to make sure that the product meets the requirements.
Learning activity concerning collaborative learning	**Communication** Students have to work together on complex tasks, so it is important that they possess some communication skills: introducing new ideas to other students, presenting their opinions to other students, explaining things to other students, listening to other students, giving other students the opportunity to explicate their ideas, etc.	Questions - How can you explain in another way what you should take into account when you build a house? - How can you make sure that everyone has the opportunity to share his/her opinions and ideas with the rest of the group? Feedback - Alerting students to the fact that everyone should be allowed to share his/her opinions and ideas. - Alerting student to the fact that they should let a person have his/her say. - Giving examples of how to introduce ideas to the group.

Appendix 1 Aspects of learning activities (Continued)

	Aspects of learning activities	Coaching interventions
Learning activity concerning collaborative learning	**Contribution to the group process and product** Students are supposed to work collaboratively on a complex task, which means that every student has to contribute to the process and to the product. This involves agreeing on what to do and keeping appointments. They also should establish an allocation of tasks that is equal for all participants, and be able to make decisions as a group. In addition, they should consult each other on work that is done, and be able to take responsibility for it.	Questions - Did you make agreements with regard to x? - Did you include the agreements in the minutes? - Dave, do you know what Pete's task is? - Susan, do you know when Dave should have finished his task? - Richard, did you call Jessica to account for not finishing her task in time? - Did you all have an equal share in completing the task? Feedback - Giving clues as to how to call a person to account for something. - Giving clues as to how to tune individual tasks to the group task.
	Contribution to group climate This learning activity involves contributing to a positive climate within the group. This requires that students respect each other, listen to each other, and independently resolve conflicts.	Questions - Do you think that this attitude will contribute to a positive group climate? - How do you think David feels right now? Feedback - Giving examples that show the importance of respecting other students in collaborative working. - Giving advice on how to solve a conflict between students.

Appendix 1 Aspects of learning activities (Continued)

	Aspects of learning activities	Coaching interventions
Affective learning activities	**Maintaining self-confidence, positive expectations, motivation, dedication & persistence, and concentration** While working on complex tasks, students experience emotions that lead to specific attitudes towards working on complex tasks (positive, neutral, negative). These emotions can influence the learning process dramatically. To develop a positive learning attitude, students need to maintain self-confidence, positive expectations, motivation, dedication & persistence, and concentration. These learning activities are an important condition for carrying out other relevant learning activities.	Questions - How do you motivate yourself to finish tasks that you don’t like? - What do you like about this task? Feedback - Complimenting students on their efforts and the results.

Appendix 2 Score form for assigning a score to a video episode

Name:………………………………. Video episode:………………………………………
Teacher judged:………………….…. Learning activity judged:…………………………….

<table>
<tr><th colspan="2">Video episode</th><th colspan="2">Interviews</th></tr>
<tr><td><b>Step 1: Collecting evidence</b>
- 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?</td><td>Score (1-4)</td><td>Teacher</td><td>Student(s)</td></tr>
<tr><td></td><td></td><td></td><td></td></tr>
<tr><td colspan="4">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.</td></tr>
<tr><td colspan="4"><b>Step 2: Assigning scores to teachers’ coaching performance</b>
Why should the assigned score be a 1, 2, 3, or 4? Write a brief summary in which you substantiate the scores assigned. In the summary, refer to or cite important arguments and evidence.</td></tr>
<tr><td colspan="4"></td></tr>
</table>

Appendix 3 Score form for assigning an overall score

Name:………………………………. Teacher to be judged:…..……………………………

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) domain knowledge and skills (b) regulation, (c) learning attitude, or (d) collaborative learning.

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.

Light (L)	Average (A)	Heavy (H)
Teachers' coaching performance had a small impact on students' growth or perception of practice. And/or The coach situation was extremely simple or complex, so that the teacher did not got the opportunity to show how well he/she can coach students.	Teachers' coaching performance had some impact on students' growth or perception of practice.	Teachers' coaching performance had a crucial impact on students' growth or perception of practice.

Overview scores assigned to separate video episodes

Video episode	1	2	3	4	5	6	7	8	9	10	Overall score
Learning activity coached											
Score and weigh	□	□	□	□	□	□	□	□	□	□	

- 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.

Why should the assigned overall score be a 1, 2, 3, or 4? 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.

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.

Appendix 4 Codebook for coding evidence and arguments

	Step 1 **Type of statements**	**Step 2** **Valence**	**Step 3** **Aspect of competent coaching**			**Step 4** **Fostered learning activity**
Consistent with the conceptual framework	Citation	Positive	**Evidence with regard to coach situation**	**Evidence with regard to teachers' behavior or appearance**	**Evidence with regard to students**	Coaching of orientation of the complex task
	Inference	Negative	Students' problem	Asking questions	Students' reaction to the interventions of the teacher	Coaching of searching and organizing relevant information
	Judgment	Neutral	Groups' problem	Providing feedback	Question to the teacher	Coaching of comprehending and using relevant subject matter
			Content of the coach session	Asking questions and providing feedback (coaching interventions)	Reaction to another student	Coaching of planning
			Aim of the teacher	Other teacher behavior	Students' learning	Coaching of monitoring
			Context factors that influence coaching	Teachers' interventions are (not) appropriate with regard to students' needs or the context	Students' thinking	Coaching of adjusting
				Missed opportunities in coaching	Students' understanding	Coaching of evaluating
					Students' growth	Coaching of motivation and dedication
					Students' awareness	Coaching of communication
						Coaching of contribution to the group process and product
						Coaching of group climate

Appendix 4 Codebook for coding evidence and arguments (Continued)

	Step 1 Type of statements	**Step 2 Valence**	**Step 3 Aspect of competent coaching**			**Step 4 Fostered learning activity**
Not consistent with the conceptual framework			Only a judgment is made (good/bad) and no arguments are reported	Only a judgment is made (good/bad) and no arguments are reported	Only a judgment is made (good/bad) and no arguments are reported	
			Learning climate	Interventions to direct the discussion		Group dynamics
				Teachers' style		Positive learning climate

Appendix 5 An example of coding evidence

Score form assessor 1		Step 1	Step 2	Step 3	Step 4
Evidence:	Judgment				
- Teacher: in the overview, sand to fill up; do we need sand?	2	Citation	Negative	Teachers' behavior (questions)	Comprehending and using relevant subject matter
- Belongs the sand to the category 'groundwork' or 'street work'?	3	Citation	Positive	Teachers' behavior (questions)	Comprehending and using relevant subject matter
- So, sand belongs to groundwork?	2/3	Citation	Neutral	Teachers' behavior (questions)	Comprehending and using relevant subject matter
- We agreed that we would cluster the activities according to categories	2/3	Citation	Neutral	Teachers' behavior (feedback)	Comprehending and using relevant subject matter
- When do we work with sand?	3	Citation	Positive	Teachers' behavior (questions)	Comprehending and using relevant subject matter
- Teacher explains the differences between groundwork and street work	-	Inference	Neutral	Teachers' behavior (feedback)	Comprehending and using relevant subject matter
- The apron is almost complete, what do I miss here?	2/3	Citation	Neutral	Teachers' behavior (questions)	Comprehending and using relevant subject matter
- (teacher asks Pete, but John is answering, the teacher asks Pete another question)	2/3	Inference	Neutral	Teachers' behavior (interventions to direct the discussion)	Group dynamics
- How do we attach the boards?	3	Citation	Positive	Teachers' behavior (questions)	Comprehending and using relevant subject matter
- How do we attach the sole?	3	Citation	Positive	Teachers' behavior (questions)	Comprehending and using relevant subject matter
- What are spring bolts?	3/4	Citation	Positive	Teachers' behavior (questions)	Comprehending and using relevant subject matter
- What do spring bolts look like?	3/4	Citation	Positive	Teachers' behavior (questions)	Comprehending and using relevant subject matter
- Is it important to know what spring bolts look like?	3	Citation	Positive	Teachers' behavior (questions)	Comprehending and using relevant subject matter

Appendix 6 An example of coding arguments

Score form assessor 1				
Summary:				
This coach session can be clearly divided in three parts: (1) ground and street activities, (2) attaching the apron, and (3) attaching the purlin.	Inference	Neutral	Coach situation (content)	Comprehending and using relevant subject matter
The teacher asks the right questions.	Judgment	Positive	Teachers’ behavior (questions)	-
And after a sequence of questions, he provides the students with a short explanation.	Inference	Neutral	Teachers’ behavior (questions & feedback)	-
He relates the domain specific knowledge to relevant situations in practice.	Inference	Neutral	Teachers’ behavior (other teacher behavior)	Comprehending and using relevant subject matter
I think that the students certainly learn from these interventions. The judgment will be a 3 or 4.	Judgment	Positive	Consequences for students (students learn)	-
The reason for assigning a 3 instead of a 4 is that the teacher provides a lot of theory to the student. I don’t think that students, who do not take notes, will remember what the teacher tries to learn them.	Judgment	Nega-tive	Teachers’ behavior (teachers’ interventions are not appropriate)	-

Nederlandse samenvatting

Hoofdstuk 1

In het eerste hoofdstuk van het proefschrift worden achtergrond, probleemstelling en onderzoeksvragen, context en relevantie van het onderzoek gepresenteerd. De ontwikkeling van instrumenten voor het beoordelen van docentcompetenties staat volop in de belangstelling. Uit onvrede met bestaande procedures, worden momenteel nieuwe beoordelingsprocedures ontwikkeld, ook wel 'authentieke performance assessments' genoemd. Een belangrijk kenmerk van deze beoordelingsprocedures is dat ze beogen recht te doen aan het complexe en contextgebonden karakter van lesgeven. In de nieuwe beoordelingsprocedures wordt veelal een mix van bewijsbronnen gebruikt die de verschillende componenten van het lesgeven bestrijken. Ook worden open taken ingezet die een beroep doen op het onmiddellijk en adequaat beslissen en handelen in de praktijk of in een context die vergelijkbaar is met de praktijk. Bij het ontwikkelen van nieuwe beoordelingsprocedures gaat ook de aandacht uit naar het waarborgen en evalueren van de kwaliteit van deze procedures. De nieuwe vormen van beoordelen brengen immers nieuwe bedreigingen van de betrouwbaarheid en validiteit met zich mee. Ten eerste spelen bij performance assessments beoordelaars een belangrijke rol; zij moeten de performance (het functioneren) van de docent interpreteren en beoordelen. Het blijkt dat het voor beoordelaars lastig is om objectief en betrouwbaar te scoren, omdat persoonlijke voorkeuren, vooroordelen en selectieve observatie moeilijk te vermijden zijn. Ten tweede wordt de validiteit van de performance assessments bedreigd door taakspecificiteit. De taken die in een assessment zijn opgenomen, blijken vaak aanzienlijk wisselende performances op te roepen bij respondenten, zelfs wanneer de taken uit eenzelfde domein komen. Ten derde is het moeilijk om een representatieve steekproef van assessmenttaken samen te stellen die alle relevante situaties en aspecten van lesgeven omvatten die docenten in de praktijk tegen kunnen komen.

Om die bedreigingen te reduceren, worden in de literatuur verschillende maatregelen aangedragen die in het design van de beoordelingsprocedure zouden kunnen worden opgenomen. Deze maatregelen bestaan uit het gebruiken van gepaste criteria en performanceniveaus, het gebruiken van een scoringsprocedure waarbij beoordelaars zorgvuldig een beoordeling opbouwen aan de hand van specifieke criteria en richtlijnen, het inzetten van meerdere beoordelaars, het gebruiken van een systematische en transparante scoringsprocedure, het trainen van beoordelaars in het toepassen van de criteria en performanceniveaus en het standaardiseren van assessmenttaken. Hoewel gaandeweg een kennisbasis ontstaat over het ontwerpen van authentieke performance assessments, blijft het relatief ingewikkeld de ontwerpprincipes uit de literatuur om te zetten in een concrete beoordelingsprocedure. In dit proefschrift wordt op basis van ontwerpprincipes uit de literatuur een performance assessment ontwikkeld en geëvalueerd. Het proefschrift levert daarmee een bijdrage aan de kennisbasis met betrekking tot het realiseren van betrouwbare, generaliseerbare en valide performance assessments.

Het performance assessment dat in dit onderzoek is ontwikkeld, werd ingezet voor het beoordelen van de coachcompetentie van docenten in het MBO, met andere woorden, de competentie van MBO-docenten in het coachen van hun leerlingen die bezig zijn met een opdracht. De beoordelingsprocedure is speciaal voor deze docentcompetentie ontworpen, omdat dit een belangrijke competentie is geworden door de implementatie van zelfstandig en competentiegericht leren in het MBO. In de context van deze innovatie is in de regio Leiden en omstreken binnen de sector Techniek het MTS+ project gestart. Binnen het MTS+ project is een leeromgeving ontwikkeld die moet bijdragen aan zelfstandig en competentiegericht leren. In deze leeromgeving is het curriculum georganiseerd rond complexe en langlopende opdrachten die sterk gerelateerd zijn aan taken die mensen tegenkomen in de beroepspraktijk. Tijdens het uitvoeren van deze opdrachten worden de leerlingen gecoacht door hun docent.

In dit onderzoek is de coachcompetentie van docenten in het MBO beoordeeld op basis van een videodossier. Een videodossier bestaat uit een mix van bewijsbronnen die een compleet overzicht geven van de coachcompetentie van een docent. De belangrijkste bewijsbronnen in het dossier zijn de videofragmenten die verschillende kritische situaties tonen waarin een docent zijn of haar coachperformance laat zien. Verder zijn vier bronnen met contextinformatie toegevoegd: een samenvatting van wat er tijdens het videofragment te zien is en van wat er vooraf ging aan het videofragment, achtergrondinformatie over de leerlingen die tijdens het videofragment te zien zijn (leeftijd, vooropleiding, begeleidingsbehoefte, enz.), een beschrijving van het lesmateriaal dat tijdens het videofragment wordt gebruikt en een interview met de docent en met de leerling(en) waarin gereflecteerd wordt op de coachsituatie. Tot slot is er een training ontwikkeld voor beoordelaars waarin centraal staat hoe de criteria voor competent coachen, de onderscheiden performanceniveaus en de scoringsregels toegepast dienen te worden tijdens het scoren van de videodossiers.

De centrale vraag van het onderzoek luidt: *in welke mate zijn beoordelingen op basis van een videodossier betrouwbaar, generaliseerbaar en valide?* Deze vraag is uitgewerkt in meer specifieke onderzoeksvragen die zijn onderzocht in drie deelstudies. De eerste deelstudie is een kleinschalig onderzoek waarin het ontwerp en de evaluatie van de beoordelingsprocedure centraal staan. In deze studie zijn de volgende onderzoeksvragen beantwoord:

1a In hoeverre komen beoordelaars tot dezelfde beoordelingen op basis van de ontworpen beoordelingsprocedure?

1b Welke aspecten van het videodossier stimuleren of belemmeren beoordelaars in het geven van valide interpretaties en beoordelingen?

In de tweede deelstudie is de betrouwbaarheid van de competentiebeoordelingen op basis van een videodossier nader onderzocht bij een grotere steekproef van beoordelaars. Daarnaast is in deze deelstudie ook de generaliseerbaarheid van competentiebeoordelingen nagegaan. De volgende onderzoeksvragen zijn beantwoord:

2a In hoeverre wordt de coachcompetentie van docenten in het MBO op basis van een videodossier betrouwbaar gescoord?

2b In hoeverre zijn de beoordelingen van afzonderlijke videofragmenten van docenten generaliseerbaar naar het beoogde universum van videofragmenten?

In de derde deelstudie zijn de betrouwbaarheid en validiteit van het scoren nader onderzocht. In deze deelstudie worden de volgende onderzoeksvragen beantwoord:

3a In hoeverre baseren verschillende beoordelaars hun beoordeling van de coachperformance in de videofragmenten op dezelfde bewijzen en argumenten?

3b Welk type bewijzen en argumenten rapporteren beoordelaars op de scoreformulieren?

3c In hoeverre rapporteren beoordelaars bewijzen en argumenten die corresponderen met het conceptuele kader dat is ontwikkeld voor het beoordelen van de coachcompetentie van docenten?

Voorafgaand aan de eerste deelstudie werd de beoordelingsprocedure ontworpen. De eerste stap in het ontwerp van deze procedure bestond uit een gedetailleerde analyse van de docentcompetentie coachen in de context van zelfstandig leren in het MBO. Op basis van deze domeinanalyse werden scoringsregels en een conceptueel kader gedefinieerd. De tweede stap in het ontwerp van de beoordelingsprocedure bestond uit de constructie van videodossiers met de hulp van een professionele filmploeg. Gedurende een periode van vier weken werden verschillende bronnen van bewijs verzameld rond een serie kritische coachsituaties in de praktijk. De derde stap in het ontwerp van de beoordelingsprocedure bestond uit het ontwerpen van een scoringsprocedure. Volgens deze procedure beoordeelden assessoren allereerst de coachperformance in afzonderlijke videofragmenten. Zij gebruikten hierbij specifieke criteria en beschrijvingen van competentieniveaus en werden aangespoord concrete bewijzen te zoeken waarop zij een beoordeling baseren. Vervolgens werd de beoordelaars gevraagd een overalloordeel te geven waarbij alle videofragmenten in beschouwing werden genomen. Ook hierbij gebruikten beoordelaars de beschrijvingen van de competentieniveaus en werden zij aangespoord hun beoordeling te onderbouwen met bewijzen en argumenten die betrekking hadden op de coachperformance in de afzonderlijke videofragmenten. Tot slot is er een training ontworpen waarin beoordelaars getraind werden in het toepassen van scoringsregels, het conceptuele kader en de scoringsprocedure. Nadat de beoordelingsprocedure was ontworpen, zijn de videodossiers beoordeeld door getrainde beoordelaars.

Hoofdstuk 2

In het tweede hoofdstuk wordt de eerste deelstudie beschreven. Het betreft een kleinschalig onderzoek waarin de interbeoordelaarsbetrouwbaarheid onderzocht werd evenals aspecten in het ontwerp van de beoordelingsprocedure die beoordelaars stimuleren of belemmeren in het maken van valide interpretaties en beoordelingen. Om een indicatie te krijgen van de overeenstemming in toegekende scores door beoordelaars, zijn scoreformulieren verzameld en analyses uitgevoerd. Op de scoreformulieren noteerden beoordelaars de scores die zij toekenden aan de getoonde performance in de videofragmenten. Voor de toegekende scores werd de Gower-coefficient bepaald als indicatie voor de interbeoordelaarsovereenstemming. Om inzicht te krijgen in aspecten van het ontwerp van de beoordelingsprocedure die beoordelaars belemmeren of stimuleren bij het komen tot valide interpretaties en beoordelingen, zijn alle beoordelaars geïnterviewd. Uit de resultaten van deze deelstudie blijkt dat op basis van de ontworpen beoordelingsprocedure een acceptabel tot hoog niveau van interbeoordelaarsovereenstemming kon worden bereikt. Beoordelaars plaatsten hierbij wel de kanttekening dat het toepassen van de scoringsprocedure een aanzienlijke hoeveelheid tijd en energie kostte. Daarnaast waren de beoordelaars van mening dat de training een noodzakelijke conditie was voor het correct toepassen van deze procedure. Verschillende aspecten van de procedure bleken beoordelaars te stimuleren bij het komen tot interpretaties en beoordelingen:

- het conceptuele kader met beschrijvingen van leeractiviteiten en gerelateerde coachinterventies hielp de beoordelaars bij het beoordelen van de relevante aspecten van een coachperformance;
- de samenvatting van wat er gebeurt tijdens een kritische coachsituatie hielp de beoordelaars de relevante aspecten van de coachperformance te beoordelen;
- de contextinformatie, vooral het interview met de docent en de deelnemer(s), hielp de beoordelaars bij het begrijpen van het handelen van de docent en de gevolgen hiervan voor de deelnemers;
- beoordelaars gaven aan dat ze ‘ongecompliceerde’ coachsituaties gemakkelijker konden begrijpen en daarom gemakkelijker konden scoren. Onder ‘ongecompliceerde coachsituaties’ werden in het algemeen coachsituaties verstaan waarbij (a) de performance van de docent overeen komt met de toelichting op de performance van de docent tijdens het interview of (b) het coachen op een specifieke leeractiviteit duidelijke te onderscheiden was van het coachen op andere leeractiviteiten of (c) de deelnemers behoefte hadden aan coaching op een duidelijk te onderscheiden leeractiviteit.

Naast de aspecten die een positieve invloed hadden op het komen tot valide interpretaties en beoordelingen, werden ook aspecten gevonden die beoordelaars daarbij belemmerden:

- de coachperformance in afzonderlijke videofragmenten bleek moeilijker te beoordelen dan de coachperformance over verschillende videofragmenten heen, omdat de afzonderlijke videofragmenten maar kleine stukjes laten zien van wat er tussen docent en deelnemer(s) plaatsvindt;
- videofragmenten die langer duurden dan 15 minuten leken niet bij te dragen aan meer valide interpretaties en beoordelingen, omdat het moeilijk was voor beoordelaars om zich langer dan 15 minuten te concentreren op de coachperformance en omdat er volgens de beoordelaars geen nieuwe essentiële informatie werd toegevoegd na 15 minuten;

- het bleek soms moeilijk om het coachen op competentieniveau twee te onderscheiden van het coachen op competentieniveau drie, wat in de beoordelingsprocedure de kritieke scheiding is tussen ‘onvoldoende’ en ‘voldoende’;
- de mate waarin een docent ‘praktijkgericht’ coacht was niet te beoordelen op basis van de ontwikkelde videodossiers omdat de docenten nauwelijks gedrag vertoonden dat in overeenstemming was met dit criterium. Daardoor konden de beoordelaars hun beoordeling van het praktijkgerichte coachen alleen baseren op negatief bewijs in termen van gemiste kansen door de docent.

Hoofdstuk 3

In het derde hoofdstuk wordt de tweede deelstudie beschreven. Op basis van verschillende analyses werd getracht een indicatie te krijgen van de betrouwbaarheid van de beoordelingsprocedure. Er werd bepaald in welke mate scoringstendenties voorkwamen in het scoren door de beoordelaars, de interbeoordelaarsovereenstemming werd vastgesteld evenals de generaliseerbaarheid van toegekende scores over beoordelaars. Deze analyses werden uitgevoerd op een grotere steekproef dan in de eerste deelstudie. Daarnaast werden verschillende analyses uitgevoerd om een indicatie te krijgen van de generaliseerbaarheid van de scores naar een universumscore. Een universumscore verwijst in dit verband naar de score die een respondent behaald zou hebben, wanneer hij of zij alle mogelijke taken zou hebben uitgevoerd die er zijn om de competentie te meten. Er werd een rangorde bepaald van videofragmenten die zeer eenduidige scores uitlokten tot videofragmenten die zeer wisselende scores uitlokten bij de verschillende beoordelaars. De videofragmenten die zeer wisselende toegekende scores uitlokten bij de verschillende beoordelaars zijn een bedreiging voor de generaliseerbaarheid. Ook werd voor elk videofragment bepaald in welke mate de toegekende scores aan de coachperformance in het fragment overeenkwamen met scores die werden toegekend aan de coachperformance in andere fragmenten. Een belangrijke conclusie van deze deelstudie is dat de designprincipes van de beoordelingsprocedure lijken bij te dragen aan betrouwbaar scoren door beoordelaars. Het gebruiken van het ontwikkelde beoordelingskader, de competentieniveaus, de scoringsvoorschriften en de ontwikkelde dossiers door de beoordelaars tijdens het scoren en het volgen van de training gaan over het algemeen samen met betrouwbaar scoren door beoordelaars. Er werd een acceptabel niveau van interbeoordelaarsovereenstemming gevonden en ook de generaliseerbaarheid van toegekende scores over de beoordelaars was hoog. Uit de resultaten bleek verder dat wanneer twee beoordelaars betrokken zijn bij de competentiebeoordeling op basis van een videodossier, een acceptabel niveau van interbeoordelaarsovereenstemming bereikt kan worden. Naast deze positieve resultaten, is uit deze deelstudie gebleken dat scoringstendenties voorkwamen. Beoordelaars waren niet in staat alle docenten even mild of streng te beoordelen. Bovendien bleken beoordelaars die een collega waren van de te beoordelen docent extreme beoordelingen te geven, zowel extreem positief als negatief.

Op basis van deze deelstudie kunnen ten aanzien van de generaliseerbaarheid over videofragmenten alleen tendensen worden beschreven. Definitieve conclusies over het minimale aantal videofragmenten dat nodig is om uitspraken te doen over het coachen van de docent kunnen dan ook niet getrokken worden. Het standaardiseren van de videofragmenten op basis van de definitie van een kritische situatie lijkt samen te gaan met positieve resultaten op het gebied van het generaliseren van toegekende scores over videofragmenten. De overeenstemming tussen toegekende scores aan een videofragment en de gemiddelde toegekende scores aan de rest van de videofragmenten is over het algemeen acceptabel tot goed, alleen de generaliseerbaarheid van toegekende scores aan de videofragmenten waarin de docent coacht op leerhouding is problematisch.

Hoofdstuk 4

In het vierde hoofdstuk wordt de derde deelstudie beschreven. In deze deelstudie is de validiteit van het scoren door beoordelaars onderzocht. Om de onderzoeksvragen van de deelstudie te kunnen beantwoorden, zijn verschillende kwantitatieve en kwalitatieve analyses uitgevoerd op de bewijzen en argumenten die beoordelaars rapporteerden op scoreformulieren om hun toegekende scores te rechtvaardigen. Op basis van deze analyses werd bepaald in welke mate constructirrelevante variantie en construct onderrepresentatie invloed hadden op het scoren door beoordelaars. Er werd een aanzienlijke variatie gevonden in bewijzen en argumenten die de verschillende beoordelaars aandroegen om een toegekende score te legitimeren. Ook wanneer eenzelfde score werd toegekend, bleken de bewijzen en argumenten uiteen te lopen. Er werd een grotere variatie gevonden in de argumenten dan in de verzamelde bewijzen. Op de scoreformulieren werd 58% tot 100% van de argumenten door maar een van de twaalf beoordelaars genoteerd. Verder bleek dat beoordelaars zowel concrete uitspraken deden over wat ze gezien hadden in het videodossier als, meer abstracte interpretaties en beoordelingen gaven van wat ze gezien hadden in het videodossier. De concrete uitspraken werden voornamelijk gebruikt bij het aandragen van bewijzen en de abstracte uitspraken voornamelijk bij het aandragen van argumenten. De concrete bewijzen hadden betrekking op het gedrag van de docent: beoordelaars noteerden de vragen en feedback die de docenten inzetten tijdens het coachen. Deze bewijzen werden beschouwd als relevante bewijzen, omdat ze pasten binnen het conceptuele kader dat de beoordelaars zouden moeten gebruiken bij het beoordelen. Beoordelaars lijken dus redelijk in staat om relevante bewijzen te identificeren. Bij de cijfermatige beoordeling, schreven de beoordelaars een toelichting waarin ze interpretaties gaven van wat ze tijdens de videofragmenten hadden gezien en ook gaven ze een waardeoordeel hierover. Beoordelaars noteerden voornamelijk argumenten die betrekking hadden op het gedrag van de docent (18%) en de coachsituatie (14%). De waardeoordelen die beoordelaars noteerden in deze toelichting hadden betrekking op het gedrag van de docent (48%) en op de consequenties van het gedrag voor de deelnemers (19%). In het algemeen waren de bewijzen en argumenten consistent met het ontwikkelde conceptuele kader dat de beoordelaars verondersteld werden te gebruiken tijdens het beoordelen en werden er weinig construct-irrelevante bewijzen en argumenten aangedragen door beoordelaars. Het scoren door beoordelaars lijkt wel beïnvloed te worden door construct onderrepresentatie. Beoordelaars waren geneigd om in plaats van alle aspecten alleen een of twee aspecten van het conceptuele kader te gebruiken bij het beoordelen van de coachperformance.

Hoofdstuk 5

Op basis van de drie deelstudies worden in hoofdstuk 5 de algemene conclusies, beperkingen, suggesties voor vervolgonderzoek en praktische implicaties van het onderzoek besproken. Op basis van de drie deelstudies zijn tien algemene conclusies geformuleerd. Vijf van deze conclusies hebben betrekking op de mate waarin beoordelingen van een videodossier *betrouwbaar* zijn:

1. beoordelaars bereikten een acceptabel tot hoog niveau van overeenstemming voor het toekennen van (overall)scores wanneer zij de coachcompetentie van docenten uit het MBO beoordeelden;
2. beoordelaars bereikten een hoger niveau van overeenstemming voor het toekennen van overallscores dan voor het toekennen van scores aan de coachperformance in afzonderlijke videofragmenten;

3. twee beoordelaars waren nodig om een acceptabel niveau van interbeoordelaarsovereenstemming te verkrijgen;
4. het scoren door beoordelaars werd beïnvloed door scoringstendenties;
5. beoordelaars baseerden hun toegekende scores op verschillende bewijzen en argumenten, waarbij meer variatie werd gevonden in argumenten dan in bewijzen.

Twee conclusies hebben betrekking op de mate waarin de beoordelingen op basis van een videodossier *generaliseerbaar* zijn:

6. de beoordelingen die werden toegekend aan de videofragmenten van bepaalde docenten waren beter te generaliseren naar het beoogde universum van videofragmenten dan de beoordelingen van andere docenten;
7. de beoordelingen die werden toegekend aan videofragmenten waarin de docent coachte op bepaalde leeractiviteiten waren beter te generaliseren naar het beoogde universum van videofragmenten dan de beoordelingen van het coachen op andere leeractiviteiten.

Tot slot zijn drie conclusies getrokken die betrekking hebben op de mate waarin de beoordelingen op basis van een videodossier *valide* zijn:

8. beoordelaars ervoeren de contextinformatie die was toegevoegd aan het videodossier als noodzakelijke achtergrondinformatie voor een valide beoordeling van de coachcompetentie van de docenten;
9. beoordelaars waren in staat om tijdens het scoren bewijzen en argumenten te gebruiken die correspondeerden met het ontwikkelde conceptuele kader;
10. de validiteit van het beoordelen op basis van een videodossier werd wellicht bedreigd door het feit dat beoordelaars tijdens het scoren maar een of twee aspecten van het conceptuele kader gebruikten in plaats van alle aspecten.

Vervolgens worden in hoofdstuk 5 enkele beperkingen van het onderzoek beschreven. Ten eerste kon als gevolg van de gebruikte steekproef in deelstudie twee geen generaliseerbaarheidstudie worden uitgevoerd, maar werd op basis van itemrest correlaties en standaarddeviaties een indicatie gegeven van de generaliseerbaarheid over videofragmenten. Ten tweede is de studie in hoofdstuk vier gebaseerd op *gerapporteerde* bewijzen en argumenten door beoordelaars op scoreformulieren. Hierdoor werden bewijzen en argumenten die beoordelaars niet rapporteerden, maar die mogelijk wel een rol speelden in het beslisproces, buiten beschouwing gelaten. Deze inperking is aangebracht, omdat het bestuderen van expliciet gerapporteerde bewijzen en argumenten een logische eerste stap is bij het onderzoeken van constructirrelevante variantie en construct onderrepresentatie in het scoren door beoordelaars. Ten derde is alleen bestudeerd *welke* bewijzen en argumenten beoordelaars aandroegen en niet *hoe* de bewijzen en argumenten gecombineerd werden tot een (eind)oordeel. Ook hier is de reden dat het bepalen van de gebruikte bewijzen en argumenten, een eerste logische stap is in het onderzoeken van het scoringsproces van beoordelaars. Pas na deze eerste stap kan onderzoek worden gedaan naar de wijze waarop beoordelaars bewijzen en argumenten combineren tot een (overall)oordeel.

Hoofdstuk 5 bevat tevens enkele suggesties voor vervolgonderzoek. De eerste lijn van vervolgonderzoek betreft het onderzoeken van de mate waarin de beoordelingen, verkregen op basis van een videodossier, kunnen worden geëxtrapoleerd naar prestaties buiten de assessmentcontext. De tweede lijn van vervolgonderzoek heeft betrekking op de mate waarin de ontwikkelde beoordelingsprocedure bijdraagt aan de professionele ontwikkeling van docenten die hebben deelgenomen aan het assessment. De derde lijn betreft vervolgonderzoek dat zich richt op het nader onderzoeken van kenmerken van videofragmenten.

Tot slot worden in hoofdstuk 5 enkele implicaties van het onderzoek beschreven voor de assessmentpraktijk. Ten eerste blijkt uit het onderzoek dat de voorgestelde designprincipes uit de literatuur die zijn toegepast bij de constructie van de videodossiers gebruikt kunnen worden voor het genereren van betrouwbare, generaliseerbare en valide beoordelingen. Over het algemeen werden positieve resultaten gevonden ten aanzien van het scoren door beoordelaars en de generaliseerbaarheid van beoordelingen wanneer de ontwikkelde beoordelingsprocedure werd ingezet tijdens het beoordelen van de coachcompetentie van docenten. Ten tweede blijkt uit het onderzoek dat in de praktijk volstaan kan worden met twee beoordelaars om tot betrouwbare beoordelingen te komen, mits het assessment wordt vormgegeven volgens de designprincipes die in dit onderzoek gebruikt zijn. Dit is een belangrijke implicatie, omdat het in de praktijk vaak niet mogelijk is om nog meer beoordelaars in te zetten bij assessments vanwege de hoge kosten die dit met zich mee zou brengen. Ten derde zijn er verschillende aanwijzingen uit het onderzoek naar voren gekomen voor de verbetering van trainingen voor beoordelaars. In dit soort trainingen zou bijvoorbeeld veel aandacht besteed moeten worden aan het onderscheid tussen performances die net wel en die net niet als voldoende kunnen worden aangemerkt. Beoordelaars blijken het moeilijk te vinden om performances op de grens van voldoende en onvoldoende te beoordelen.

Daarnaast blijkt dat beoordelaars tijdens de training gestimuleerd moeten worden om alle aspecten van het conceptuele kader te gebruiken, omdat ze anders geneigd zijn sommige aspecten buiten beschouwing te laten tijdens het beoordelen. Een aspect dat ook expliciet in de training tot uitdrukking zou moeten komen, is het creëren van een gedeeld conceptueel kader, zodat alle beoordelaars hetzelfde verstaan onder de competentie die ze moeten beoordelen en het conceptuele kader dat ze gebruiken tijdens het beoordelen. Deze maatregel zou de aanzienlijke variatie in aangedragen bewijzen en argumenten moeten verminderen. De laatste aanwijzing voor assessorentrainingen bestaat eruit dat er tijdens de training intensief geoefend moet worden in het bepalen van de consequenties van het handelen (coachen) van de docent voor de deelnemers. Beoordelaars waren erop gericht om concreet gedrag van de docent te beoordelen en waren minder geneigd om ook de consequenties van het gedrag voor de deelnemers mee te nemen in de beoordeling.

Publications

Scientific publications

Bakker, M., Sanders, P., Beijaard, D., Roelofs, E., Tigelaar, D., & Verloop, N. (2008). De betrouwbaarheid en generaliseerbaarheid van competentiebeoordelingen op basis van een videodossier. *Pedagogische Studiën 85*(4), 240-260.

Submissions

Bakker, M., Roelofs, E., Beijaard, D., Sanders, P., Tigelaar, D., & Verloop, N. (submitted). Video portfolios: The development and practical utility of an authentic teacher assessment procedure.

Bakker, M., Sanders P., Beijaard, D., Roelofs, E., Tigelaar, D., & Verloop, N. (submitted). Reliability and generalizability of performance judgments based on a video portfolio.

Bakker M., Beijaard, D., Roelofs, E., Tigelaar, D., Sanders, P., & Verloop, N. (submitted). The impact of construct-irrelevant variance and construct under-representation in assessing teachers' coaching competence.

Papers

Bakker, M., Beijaard, D., Roelofs, E., Sanders, P., & Verloop N. (2005). *Video-based assessment of teachers' coaching competence.* Paper presented at the ISATT conference, Sydney, Australia.

Bakker, M., Beijaard, D., & Roelofs, E. (2005). *Het beoordelen van docentcompetentie op basis van een videodossier: implicaties voor het verzamelen van bewijsmateriaal.* Paper presented at Onderwijs Research Dagen 2005, Gent, Belgium.

Bakker, M., Beijaard, D., Roelofs, E., Sanders, P., Tigelaar, D., & Verloop N. (2007). *De generaliseerbaarheid van competentiebeoordelingen bij docenten op basis van een videodossier.* Paper presented at Onderwijs Research Dagen 2007, Groningen, The Netherlands.

Bakker, M., Beijaard, D., Roelofs, E., Tigelaar, D., Sanders, P., & Verloop, N. (2007). *Video portfolios: The development and practical utility of an authentic teacher assessment procedure.* Paper presented at a VOR division conference, Utrecht, The Netherlands.

Poster

Bakker, M., Beijaard, D., & Roelofs, E. (2004). *Video-based assessment of teachers' coaching competence.* Poster presented at the Second Biannual Joint Northumbria/EARLI SIG Assessment Conference, Bergen, Norway.

Other publications

Bakker, M., Roelofs, E., & Beijaard, D. (2006). Docentbekwaamheid in beeld gebracht met videodossiers. In E. Roelofs & G. Straetmans (Eds.), *Assessment in actie: Competentiebeoordelingen in opleiding en beroep* (pp.163-190). Arnhem: Cito.

Curriculum Vitae

Mirjam Bakker was born on the 31th of March 1977, in Beverwijk, the Netherlands. She attended secondary education at the Augustinus College in Beverwijk, where she graduated in 1996. From 1996 to 2001 she studied educational science at the University of Amsterdam and graduated cum laude. Her master's thesis concerned the effects of student characteristics on studying texts. After her graduation, Mirjam worked as a junior-publisher for Swets Test Publishers in Lisse. From 2001 to 2003, she coordinated the publishing of tests such as IQ-tests, spelling tests, arithmetic tests, and tests for career advisory which can be used in the field of education. In 2003, Mirjam started as a PhD student on a research project at ICLON Graduate School of Teaching at Leiden University and Cito. Her research focused on reliability and validity issues in authentic performance assessments for teachers. Currently, Mirjam works at the ICLON Graduate School of Teaching at Leiden University. At present, she is involved part time in a postdoctoral research project on the impact of different formative assessment approaches on teachers' competence development and part time in teacher education.

Dankwoord

Op deze plek in het proefschrift wil ik van de gelegenheid gebruik maken om verschillende mensen te bedanken die een rol hebben gespeeld in mijn promotietraject.

Allereerst de docenten en het management van ROC Leiden die betrokken zijn geweest bij de ontwikkeling van de videodossiers. Ronald Stam, Hans van Ballengooij, Hannie van Beelen, Andy Hoogendoorn en Jan Berkhout: heel hartelijk bedankt voor alle inspanningen, niets was teveel voor jullie! Ook alle andere docenten van de locatie DPL wil ik bedanken voor het aangename verblijf op het ROC. Wat hebben we gelachen: de verhalen over de sportdag in de polder, de koffie met zout, de bekraste cd met filmmateriaal en het bezoek van Sinterklaas op mijn verjaardag zullen me nog lang bijblijven. Wat een voorrecht om op deze manier onderzoek te doen! Ook wil ik speciaal bedanken de docenten die zich door de videodossiers hebben geploegd om deze te beoordelen. Wat was het veel werk! Jos Atteveld, Hubert Heuzen, Jan Otto, Ronald Hanselaar, Cees de Rooij, Christel Matla (ROC Leiden), Cosiene Burger, Wilma van Raaij, Rob Aartsen (ROC Amsterdam Gooi en Vechtstreek), Martijn Groen, Erik Deuling, Marija Westerlaken (Nova College): heel erg bedankt, zonder jullie inspanningen had ik niet zoveel en zulke rijke onderzoeksdata gehad.

Alle collega's in het land en op het ICLON: het was fijn om met jullie samen te werken. Speciaal wil ik de onderzoeksgroep van het ICLON bedanken die tijdens mijn promotietraject op verschillende momenten heeft meegedacht over mijn onderzoek. Bedankt voor jullie inspirerende en constructieve ideeën. Ook mijn collega-aio's wil ik bedanken voor de steun en gezelligheid die jullie me hebben gegeven. Met veel plezier denk ik terug aan de vele lunchpauzes, aio-etentjes, cursussen, en congressen. In het bijzonder wil ik mijn kamergenoten Mirjam, Ineke en Christel bedanken. Mirjam, bedankt voor de vele nuttige discussies op het gebied van assessment en je droge humor! Ineke, dankjewel voor de wijze adviezen en de gezellige gesprekken. Christel, jij hebt de laatste maanden van mijn promotietraject meegemaakt, dankjewel voor je aanstekelijke enthousiasme!

Ook de mensen van het Cito wil ik bedanken voor de warme ontvangst dat ik heb gekregen wanneer ik langs kwam. In het bijzonder wil ik Chris Phielix, Ton Heuvelmans en Piet Sanders bedanken. Chris, jouw hulp bij het samenstellen van de videodossiers was onmisbaar, je hebt heel veel en goed werk verzet. Ton, bedankt voor je ondersteuning bij de analyses van de tweede deelstudie. Piet, bedankt voor het meedenken over de opzet van het onderzoek, de analyses en het schrijven van de artikelen, ik heb er heel veel van geleerd!

Tot slot wil ik mijn vrienden en familie bedanken en in het bijzonder Frans, Riet, Jeroen, Marieke, Martijn, en Anoeskha. Jullie vierden de successen met me mee en steunden me als het minder ging. Heel erg bedankt daarvoor. Lieve Michel, halverwege mijn promotietraject kwam jij voorbij. Vlak voor de afronding van het proefschrift hebben we samen een huis gekocht. Het was erg druk. Jij hebt veel dingen uit handen genomen, zodat ik het proefschrift kon afmaken. Dankjewel voor je geduld en je waardering.

Mirjam Bakker, oktober 2008

ICLON

Leiden University Graduate School of Teaching

PhD dissertation series

Hoeflaak, A. (1994). *Decoderen en interpreteren: een onderzoek naar het gebruik van strategieën bij het beluisteren van Franse nieuwsteksten.*

Verhoeven, P. (1997). *Tekstbegrip in het onderwijs klassieke talen.*

Meijer, P.C. (1999). *Teachers' practical knowledge: Teaching reading comprehension in secondary education.*

Zanting, A. (2001). *Mining the mentor's mind: The elicitation of mentor teachers' practical knowledge by prospective teachers.*

Uhlenbeck, A.M. (2002). *The development of an assessment procedure for beginning teachers of English as a foreign language.*

Oolbekkink-Marchand, H.W. (2006). *Teachers' perspectives on self-regulated learning: An exploratory study in secondary and university education.*

Henze, F.A. (2006). *Science teachers' knowledge development in the context of educational innovation.*

Mansvelder-Longayroux, D.D. (2006). *The learning portfolio as a tool for stimulating reflection by student teachers.*

Meirink, J.A. (2007). *Individual teacher learning in a context of collaboration in teams.*

Nijveldt, M. (2007). *Validity in teacher assessment: An exploration of the judgement processes of assessors.*

Bakker, M. (2008). *Design and evaluation of video portfolios: Reliability, generalizability, and validity of an authentic performance assessment for teachers.*