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

Bakker, M.E.J.

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.