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

Bakker, M.E.J.

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 κ) 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 assesssors	Similar citations reported by 5-8 assesssors	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 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.

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.