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

Bakker, M.E.J.

Chapter 5

General conclusions and discussion

5.1 Overview of the study

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

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

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

Second, in order to get an indication of the interrater agreement, assigned scores to video episodes and assigned overall scores were collected and the interrater agreement was determined. Furthermore, a semi-structured interview was carried out with all assessors in order to obtain information concerning aspects of the assessment procedure that stimulated or hindered assessors in making valid interpretations and judgments. The main findings of this study were that an acceptable to high level of agreement between assessors was found which indicates that assessors arrived at corresponding scores. However, assessors indicated that mastering the scoring

procedure takes time and energy. They perceived the assessor training as a necessary condition for applying the scoring procedure in the right way. Several factors were found to be helpful for assessors in making valid interpretations and judgments. Assessors indicated that the following factors assisted them in making valid interpretations and judgments:

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

Some disabling factors were also indicated by assessors, namely:

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

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

In study 3, the validity of assessors' scoring process was investigated. A qualitative content analysis was conducted on evidence and arguments that assessors reported on score forms to justify their assigned scores. Based on this analysis, the impact of

construct-irrelevant variance and construct under-representation of assessors' judgments was examined. A considerable amount of variation was found in the reported evidence and arguments. Furthermore, more variation was found in arguments than in evidence. Assessors used a mix of concrete and abstract statements; concrete statements were predominantly used as evidence and abstract statements predominantly as arguments. The evidence and arguments were consistent with the conceptual framework, so that little construct-irrelevant evidence and arguments were reported by the assessors. However, the assessors scoring seemed to be influenced by construct under-representation, because of their tendency to focus on only one or two aspects of the conceptual framework when interpreting and judging video episodes, instead of all aspects.

5.2 Conclusions and discussion

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

5.2.1 Reliability of judgments based on a video portfolio

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

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

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

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

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

video episodes can sometimes have a too fragmented character. Assessors indicated that based on five to six video episodes they could get a pretty clear view of teachers' performance.

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

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

Conclusion 4: Assessors' scoring showed scoring tendencies.

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

show that some assessors who judged their colleagues also assigned more severe judgments. This cannot be explained by known scoring tendencies from literature and might have to do with personal characteristics of the assessor(s).

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

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

From the results of study 3 it appears that variation between assessors especially arises in formulating (abstract) arguments. This result is in line with results found by Schutz and Moss (2004), who concluded that assessors can make very different, but legitimate interpretations based on the same evidence when judging portfolios. This finding might be explained by the fact that especially in interpreting observations a system of constructs is involved in which (personal) associative connections exist (Carlston, 1992, 1994; DeNisi, Cafferty, & Meglino, 1984; Feldman, 1981; Landy & Farr, 1980)

and which influence the assessors' evaluation of observations. Considering the variation that was found in arguments, the assessors seemed not to evaluate teachers' coaching performance based on a totally shared understanding of constructs and associations concerning competent coaching. It might be that more training sessions are needed to establish a shared system of constructs, than the four sessions that were used in this study.

5.2.2 Generalizability of judgments based on a video portfolio

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

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

The results of study 2 show that the generalizability of scores assigned to video episodes of teacher 1 and teacher 2 were better generalizable to a universe of video episodes than scores assigned to video episodes of teacher 3 and 4. The generalizability of scores assigned to video episodes of teacher 3 was the lowest. Based on the results of study 2, it is hard to predict the reason why the scores assigned to video episodes of teacher 1 and 2 could be better generalized than the scores assigned to video episodes of teacher 3. It might be that the teachers 1 and 2 reacted more consistent to the coaching situations than teacher 3. But it might also be that the assessors scored teacher 1 and 2 in a more consistent way than teacher 3. In study 2, the lowest level of interrater agreement was found for teacher 3. This result shows that also in study 2, problems with the scoring of the video portfolio of teacher 3 were detected. Furthermore, assessors reported in study 1 that especially video episodes that were longer than 15 minutes were hard to score. The video episodes that were included in the video portfolio of teacher 3 were predominantly longer than 15 minutes, which might have influenced the assessors' scoring. In study 1, assessors also

reported that complex video episodes were hard to score. They indicated that especially video episodes in which students needed support in multiple learning activities at once and where the teacher was coaching on several learning activities at the same time were hard to score. The students in the video episodes of teacher 3 had severe motivation problems (which pertain to affective learning activities), which also led to problems with regard to collaborative processes (which pertains to learning activities with regard to collaborative learning). These two problems were present in all video episodes of teacher 3 and the teacher was expected to address these hard problems. It seemed that the video episodes of teacher 3 are typical examples of what the assessors indicated as 'complex video episodes', a factor thus that influenced the scoring of the video portfolio of teacher 3. These findings suggest that in order to be able to generalize scores to a universe, it is recommended to include video episodes in the video portfolios that are less complex and last no longer than 15 minutes. Sometimes complex video episodes cannot be avoided. In that case, more video episodes could be included in a video portfolio to realize generalizability or more assessors could be used to judge the complex video episodes.

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

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

generalizability of scores assigned to video episodes in which the teacher coaches on affective learning activities. It is expected that especially the inclusion of more video episodes with regard to the coaching of these learning activities will improve the generalizability.

5.2.3 Validity of judgments based on a video portfolio

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

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

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

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

Although a lot of variation was found between assessors with regard to evidence and arguments (conclusion 5), the variation seemed not to be caused by the use of irrelevant evidence and arguments. Most of the scoring process was based on

construct relevant evidence and arguments. This was also found in other studies (Heller, Sheingold, & Myford, 1998; Nijveldt, 2007). Only 1% of the evidence and 4% of the arguments were irrelevant compared to the conceptual framework. In addition, little more construct-irrelevant evidence and arguments were found in assessors' judging of the coaching on a specific learning activity. It seemed that assessors had trouble with judging the coaching on a specific learning activity and, when doing that, to exclude judging of the coaching on other learning activities. A plausible explanation for the use of evidence and arguments that are related to the coaching on other learning activities is that, in practice, the coaching of different types of learning activities are so interwoven and interconnected that it is hard for assessors to judge only the coaching on a single learning activity. This explanation implies that a strict distinction between several types of learning activities is for assessors less useful in practice. However, it could also be that assessors just needed more training in judging the coaching on a specific learning activity in order to be able to identify evidence and use arguments that are related to the coaching of the learning activity that assessors were expected to judge.

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

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

coaching may explain the large variation that was found in reported evidence and arguments on score forms.

5.3 Limitations of the study

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

Size of the sample of video episodes

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

Focus on reported evidence and arguments on score forms

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

representation on the validity of assessors' scoring might have been larger than was found in study 3. After all, the construct-irrelevant variance and construct under-representation for evidence and arguments that assessors used, but not wrote down on the score forms, were not covered in this study.

Combining evidence and arguments to a judgment

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

5.4 Suggestions for future research

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

Extrapolation to performance outside the assessment context

The aim of the studies presented in this dissertation was to investigate the internal validity of the performance assessment (Lissitz & Samuelson, 2007); the focus was on assessors' scoring and the generalization of scores to a universe of scores. However, another vital aspect of validity is the relation between assessment scores and external measures (extrapolation inference; Kane, 2006). In the design of the video portfolios, several measures were taken to warrant the extrapolation to performance outside the assessment context: (1) high-fidelity assessment tasks were used that represent the complex situations that teachers face in practice, (2) domain coverage was expected to be realized by including ten video episodes in each video portfolio that covered the coaching in different learning activities, and (3) the video portfolios encompassed four

weeks in which students worked on one complex task. However, the contribution of these design principles to the possibility of extrapolation of scores was not investigated in this study and might be the topic of future research. This type of research includes a job analysis that shows what situations teachers face in practice and how often. Subsequently, the sample of assessment tasks should be tuned to this job analysis in order to realize content coverage. Based on this type of research, the design principles to ensure extrapolation can be adjusted and refined in order to further improve the (methodological) quality of performance assessments.

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

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

Characteristics of the video episodes

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

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

5.5 Implications for assessment practices

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

Design of the assessment procedure

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

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

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

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

Assessors

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

Assessor training

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

Final Remark: practical feasibility of video portfolios

The video portfolios designed in this study, were primarily constructed in order to investigate proposed design principles in the literature and in order to obtain new insights in processes and factors that affects the reliability, generalizability, and validity of performance assessments such as video portfolios. The practical feasibility of the

video portfolios had less priority in the design of the video portfolios. The idea behind this approach was to investigate the reliability, generalizability, and validity under 'ideal conditions'. The assumption was that when the methodological quality of the assessment could not be ensured under ideal conditions, that it will be impossible to realize this in practice.

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

References

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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