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

Bakker, M.E.J.

Chapter 3

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

Abstract

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

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

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

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

3.1 Introduction

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

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

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

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

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

Several design principles can be used that can ensure the tenability of the scoring, generalization, and extrapolation inference. Examples of such design principles are increasing the number of assessors, standardising assessment tasks, and using

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

3.2 Validity and reliability in scoring, generalization, and extrapolation

Reliability is defined as the extent to which the results of an assessment can be repeated. It entails the question of whether assessment results will vary when the assessment is repeated under the same conditions. In recent decades, the definition of validity has undergone some changes. Three perspectives on validity have been distinguished: criterion validity, content validity, and construct validity. Criterion validity refers to the relationship between the test score and an external criterion that is viewed as a direct measurement of the characteristic to be measured. Content validity refers to the extent to which the measurement is representative of the domain to be measured. Construct validity concerns the extent to which the construct (or characteristic) to be measured, is measured. Nowadays, this traditional classification of validity receives less support. Construct validity is now seen as a term that also covers

criterion validity and content validity (Messick, 1989). Validity is seen as "an integrated evaluative judgment of the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of inferences and actions based on test scores or other modes of assessment" (Messick, 1989, p.13). Although some objections can be made against this definition of validity, like that it is very broad (Borsboom & Mellenbergh, 2004), it has been generally accepted since the eighties.

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

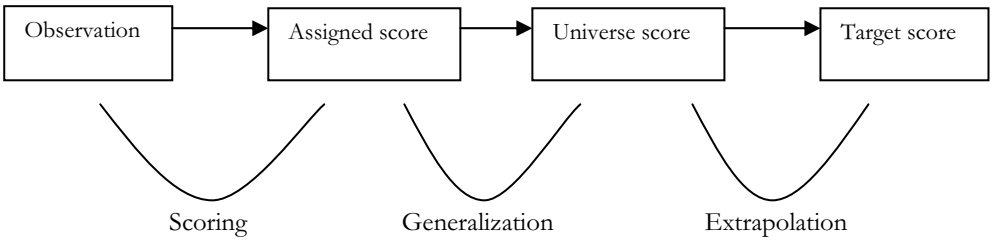


Figure 3.1 Chain of inferences for a validity argument regarding performance assessments

Scoring of performance

The first inference from the chain pertains to the scoring of the performance of respondents by the assessor: are the assessor's interpretations and judgments of the performance valid and reliable? Especially the influence of personal characteristics on judgments is a serious threat to the tenability of the first inference regarding scoring, such as selective observation, biases, and personal prejudices (Gipps, 1994; Moss, 1994). Several factors can influence the tenability of the scoring inference. First, judgments are more valid and reliable when appropriate criteria, performance levels, and scoring rules are used during the scoring process and when assessors are capable of applying these in a consistent way. Assessor training has a positive influence on the application of criteria, standards, and scoring rules (Day & Sulsky, 1995; Stamoulis & Hauenstein, 1993). Second, a large number of assessors contributes generally to more reliable scoring (Kane, 2006). When multiple assessors judge a performance, the

personal influence on the judgment of individual assessors decreases, so that the scores assigned are more accurate. Third, it appears that assessors score a performance in a more consistent way when all respondents perform the same assessment tasks instead of different tasks. This mainly leads to more reliable judgments (Crooks, Kane, & Cohen, 1996).

Generalization across assessment tasks

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

Extrapolation to performance outside the assessment context

In examining the third inference, it is investigated to what extent it is possible to extrapolate the performance as measured in the assessment to performance outside the assessment context. A design principle used to enable extrapolation to performance outside the assessment context is the use of so-called 'high-fidelity tasks' (Kane, 2006). These tasks measure the characteristic in a very direct way. However, high-fidelity tasks are often complex and open-ended tasks that are hard for assessors to score. Furthermore, these tasks are very time consuming, so that for reasons of

practical feasibility, only a restricted number of tasks can be included in an assessment. As a result of the restricted number of assessment tasks, it can be hard to establish a representative sample to enable extrapolation to performance outside the assessment context. Especially the use of a large number of assessment tasks has a considerable positive effect on extrapolation to performance outside the assessment context (Dunbar, Koretz, & Hoover, 1991; Ruiz-Primo, Baxter, & Shavelson, 1993). This remains a difficult issue in performance assessment; no clear-cut solution is at hand.

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

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

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

3.3 Method

3.3.1 Design of the performance assessment procedure

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

Following this, an assessment scale was constructed to enable expression of the level of performance. The starting point in constructing the performance levels was the definition of competent teaching by Roelofs and Sanders (2007). They see competent teaching as being able to make appropriate and deliberate decisions in a specific context, based on a personal knowledge base, which results in behavior that contributes to desired consequences. Competent coaching was then defined. In this study, competent coaching was defined as constructive coaching. Constructive coaching entails that the teacher uses coaching interventions that provide students with opportunities and stimulate them to improve the self-regulating learning activities described above. In constructive coaching, the teacher provides just enough support so that the students can make the step to a higher level in employing learning

activities, which they couldn't have made on their own (Vygotsky, 1978). As the performance of a learning activity improves, the support of the teacher decreases until the student can perform the learning activity by him/herself; this is referred to in the literature as 'fading' (Collins, Brown, & Newman, 1989). Table 2.2 in chapter 2 presents the performance levels of (non-) constructive.

Video portfolios

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

Scoring procedure

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

3.3.2 Measures to achieve reliable and valid scoring

Scoring guide and related conceptual framework

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

Theory and practice

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

Concrete examples of coaching interventions

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

Use of performance levels

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

Scoring procedure

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

Assessor training

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

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

- identifying, selecting, and quoting evidence from video episodes which is/is not consistent with the conceptual framework;
- evaluating evidence and reasoning about evidence in terms which are/are not consistent with the conceptual framework;
- assigning scores to video episodes which are/are not based on the designed performance levels for constructive coaching;
- evaluating performance across video episodes and reasoning about performance across video episodes in terms that are/are not consistent with the conceptual framework;
- assigning scores to the complete video portfolio which are/are not consistent with the conceptual framework.
- writing a rationale in which assigned scores are legitimized.

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

Organization and arrangement of evidence

In order to ensure validity and reliability in assessors' interpretations and judgments, three measures were taken. First, a professional video production company recorded the videos. Three cameras and three microphones were used to record all teacher and student activities at the same time. The starting point was that all interactions between

teacher and student(s) would be clearly perceptible for assessors, to ensure that no evidence would be lost. Second, in addition to video episodes, supporting sources of information that outlined the coaching context were included in the video portfolios. It appears that assessors need this information to be able to decide on the level of the coaching performance shown in the video episodes (Heller, Sheingold, & Myfords, 1998; Schutz, & Moss, 2004). Third, the video episodes and context information were visually ordered in a multi-media environment, to enable assessors to evaluate all available evidence in coherence.

3.3.3 Measure to generalize across video episodes

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

3.3.4 Measures to extrapolate to performance outside the assessment context

As mentioned earlier in this study, the measures applied in the assessment procedure in order to extrapolate to performance outside the assessment context were not evaluated in this study. Nevertheless, the measures applied are described in this section. In the video portfolio performance assessment, high-fidelity tasks were used to measure teachers' coaching competence in a very direct way. The high-fidelity tasks were actual coaching tasks that teachers carried out in their classrooms, as a result of emerging learning needs on the part of the students. From all recordings made in the classroom, key situations were selected for inclusion in the video portfolio. In order to be able to extrapolate to teachers' coaching competence outside the assessment context, it was important to create a sample of coaching situations that represented coaching situations that would occur in practice. To establish variation in the video episodes, the video episodes of different key situations were selected on the basis of the following criteria: the sample should contain key situations spread across the four

weeks that students worked on one complex task, and covering all stages of learning that might take place. In addition, the sample should contain video episodes that concerned coaching in all the different learning activities. Another important factor in creating a sample of video episodes is the number of video episodes to be included in the video portfolio. The larger the number of video episodes included in the portfolio, the better can be extrapolated to coaching competence outside the assessment context. However, practical feasibility also plays a role here. Assessors can only score a restricted number of video episodes within a reasonable amount of time. Thus, an important consideration is how many video episodes should be included in order to be able to extrapolate, which can also be scored within a reasonable amount of time. In this study, we included ten video episodes in a video portfolio.

3.3.5 Participants

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

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

3.3.6 Data collection

After the four training sessions, the assessors scored the four video portfolios independently. They assigned a score for constructive coaching to the coaching performance in each video episode, corresponding to one of the four levels of

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

3.3.7 Analysis: Assessors' scoring

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

In a second analysis the interrater agreement on assigned scores was examined. In this type of analysis it is common to exclude the assessors who assigned the most extreme scores. For that reason the analyses were conducted twice: once including the extreme assessors and once excluding them. In this study, the frequency of cases where 50% or more of the assessors assigned the exact same (overall) score was used as an indication of agreement. The Gower coefficient was also used as an indication of interrater agreement with regard to assigned (overall) scores. A generalizability coefficient is usually used as an indicator for rater agreement. Variance components of respondents, assessors, assessment tasks, and interaction effects between these facets are estimated

in a generalizability study. However, owing to the small variation in the assigned scores found in this study, a generalizability coefficient could not be used as an indicator of interrater agreement.

The Gower coefficient is based on absolute differences between assigned scores. In addition, the range of the assessment scale is taken into account. The coefficient is not only based on the cases where assessors assign the exact same score to a performance, but also takes into account the absolute distance between the assigned scores on the assessment scale when assessors do not assign the same score.

The formula for determining a Gower coefficient is the following:

$$G_{xy} = 1 - \{ \Sigma \mid Xi - Yi \mid / nR \}$$

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

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

3.3.8 Analysis: Generalization across video episodes

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

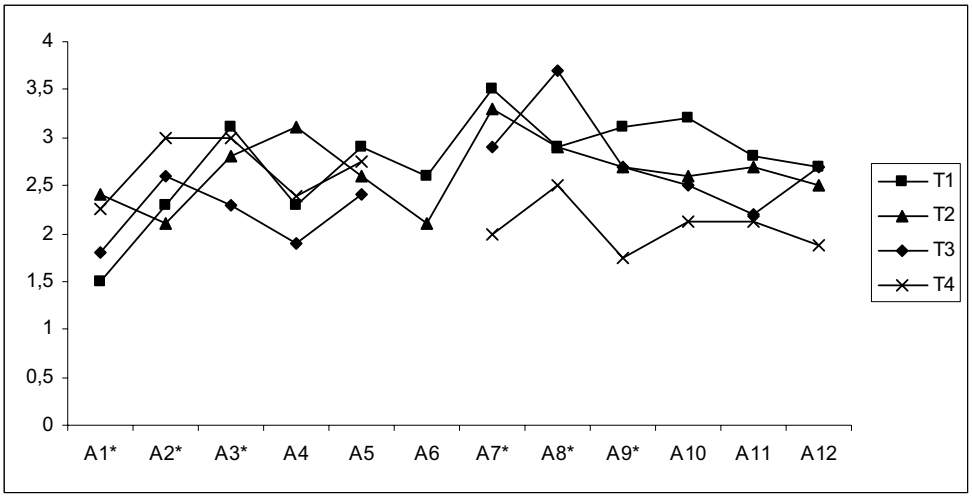
In a second analysis it was determined to what extent a score assigned to a specific video episode matched the scores assigned to other video episodes of the same type. The agreement in assigned scores to the video episodes was used to obtain an indication of the generalizability to the universe of video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

3.4 Results

3.4.1 Assessors' scoring

Tendencies in scores assigned by assessors

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



* These assessors were colleagues of the teachers assessed

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

Based on the findings of these analyses, it appears that mainly colleagues of the teachers judged assigned extreme scores. Figure 3.2 shows that assessor one gave the most severe judgment to teacher one. Assessors two and six assigned the most severe judgment to teacher two; assessor one to teacher three; and assessor nine was the

most severe assessor for teacher four. Figure 3.2 also allows the most lenient assessors for each teacher to be determined. Subsequently, it was determined which assessors assigned extreme scores. In 90% of the cases, an extreme score was assigned by an assessor who was a colleague of the teachers assessed. In 60% of the cases, an extreme overall score was assigned by an assessor who was a colleague of the teachers assessed.

Interrater agreement: Frequency

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

Interrater agreement: Gower coefficient

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

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

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

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

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

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

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

Generalizability across assessors

The interrater agreement for the average score between two assessors and the average score across ten assessors appeared to be 0.88 to 0.91. These results indicate that the average score based on ten assessors can be estimated quite accurately based on the

average score between two assessors. When the extreme assessors were included in the analysis, Gower coefficients were found to be between 0.72 to 0.90 for the average score across two assessors and across twelve assessors. Even when extreme assessors were included, an acceptable to high level of consistency was found for average scores across two and twelve assessors.

3.4.2 Generalization across video episodes

Interrater agreement for specific video episodes

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

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

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

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

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

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

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

3.5 Conclusion and discussion

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

Assessors' scoring

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

about the teacher that is relevant to the judgment of the teacher's coaching competence (Schutz & Moss, 2004). It is also possible, however, that in judging their colleagues, assessors are influenced by their biases and expectations concerning the colleagues, despite the highly structured scoring procedure.

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

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

Generalization across video episodes

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

Only tendencies with regard to the generalizability of scores across video episodes can be described on the basis of the results of this study. No conclusions can be drawn

regarding the minimum number of video episodes needed to establish an acceptable level of validity. The standardization of video episodes based on a definition for key situations appeared to go together with predominantly positive effects on generalizability. The agreement on scores assigned to a specific video episode and the average score assigned to other video episodes of the same assessment scale is predominantly acceptable to high; only the agreement on video episodes concerning coaching in affective learning activities is problematic.

Extrapolation to performance outside the assessment context

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

Future research

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

video episodes is needed. When more video episodes are scored, a generalizability study can be done on the scores assigned. These analyses reveal how much variance can be attributed to the different aspects of a performance assessment (assessors, tasks, person, and interaction effects). Furthermore, based on the findings of these analyses, conclusions can be drawn about the number of video episodes needed for a valid assessment.

References

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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