ELSEVIER

Studies in Educational Evaluation 31 (2005) 27-55

Studies in Educational Evaluation

www.elsevier.com/stueduc

COGNITIVE REPRESENTATIONS IN RATERS' ASSESSMENT OF TEACHER PORTFOLIOS

Marieke F. van der Schaaf*, Karel M. Stokking* and Nico Verloop**

**Utrecht University, Department of Educational Sciences, the Netherlands*
***Leiden University, ICLON Graduate School of Education, the Netherlands*

Abstract

Portfolios are frequently used to assess teachers' competences. In portfolio assessment, the issue of rater reliability is a notorious problem. To improve the quality of assessments insight into raters' judgment processes is crucial. Using a mixed quantitative and qualitative approach we studied cognitive processes underlying raters' judgments and the reliability of these judgments. Six raters systematically assessed 18 portfolios. The interrater reliability of 12 portfolios was satisfactory. Variance analysis showed slight rater effects. We used the Correspondent Inference Theory (Jones & Davis, 1965) and the Associative Systems Theory (Carlston, 1992; 1994) to analyse judgment forms and retrospective verbal protocols. Raters' cognitive representations on the dimensions *abstract–concrete* and *positive–negative* were significantly related to the judgments given and to the reliability of these judgments.

Introduction

In many countries interest is growing in the assessment of teachers' competences (knowledge, skills, and attitudes). This interest is fostered by current needs for

0191-491X/05/$ - see front matter © 2005 Published by Elsevier Science Ltd.
doi: 10.1016/j.stueduc.2005.02.005

accountability and quality improvement in the teaching profession (Cochran-Smith & Fries, 2001). Paralleling the movement towards alternative assessment of students (Boud, 1990; DeCorte, 1996; Dierick & Dochy, 2001), teachers' competences, too, are increasingly assessed by performance assessments, using instruments such as portfolios (Delandshere & Petrosky, 1994; Delandshere & Arens, 2001). Depending on its content and form, a portfolio can do justice to the fact that teaching is a complex activity and that teachers' behaviour is bound up with their cognitions and the teaching context (Andrews & Barnes, 1990; Bird, 1990; Lyons, 1998).

The increased use of performance assessment prompts discussions about the appropriateness of traditional psychometric criteria, such as reliability and validity, for such new forms of assessment. Reliability is often seen as a prerequisite for validity, but attaining acceptable levels of reliability in portfolio assessments is a notorious problem. This is due to the fact that assessing portfolios involves complex interactions between teachers' competences, the portfolio, the standards used, raters' characteristics, and raters' interpretations. Several studies have shown that in portfolio assessment, raters' cognitive activities are a major source of variance (DeNisi, Cafferty & Meglino, 1984; Feldman, 1981; Landy & Farr, 1980). Those cognitive activities can be described from the point of view of the content of a rater's cognitions (e.g., focussing on a teacher's behaviour or on the rater's personal beliefs), their type (e.g., interpretation or judgment), and their character (e.g., more or less explicated, situated and personal). In this study we focus on the content and the type of raters' cognitive activities in assessing teaching portfolios, especially raters' cognitive representations.

To describe raters' cognitive activities while assessing other people several general models can be used (Gilbert, 1989; Jones & Davis, 1965). However, raters' cognitive representations in assessing teacher portfolios are rarely examined. Insight into these representations is needed to facilitate and enhance our understanding of portfolio assessments and to improve their quality. This study is focused on three questions:

1. What is the reliability of teacher portfolio assessments?
2. How can raters' cognitive representations during assessment be described?
3. What is the relationship between raters' cognitive representations, their judgments and the reliability of these judgments?

The study is part of a larger research project on the assessment of experienced teachers' competences. In two previous studies we developed both content standards (criteria) and performance standards (cut-off scores) for teaching research skills in upper-secondary (ages 16–18) social science education (Van der Schaaf, Stokking, & Verloop, 2003, 2004). Teaching research skills is a recent Dutch curricular innovation which fits well with the change in many countries towards more constructivist views on learning and developing students' skills in an active, self-directed and collaborative way.

The Reliability and Validity of Raters' Judgments

Psychometric Quality Criteria

Portfolio assessment is a tool for ascertaining whether teachers satisfy required competences and for generating guidelines for professional development. In the first case, summative assessment is used to account for the teacher's expertise, with possible consequences such as certification and merit pay. In the second case, assessment has a formative goal and produces information that can be used for planning further professional development. In this article, we focus on the assessment of experienced teachers for the purposes of both summative and formative assessment.

Portfolio assessments should meet certain quality criteria. Main quality criteria are reliability and validity. Reliability refers to the extent of stability between measurements and between raters. If the same rater repeats the assessment after an interval of several weeks, or if another rater also assesses the portfolios, the judgments should not differ too much. Validity refers to the question whether the assessment is measuring the intended construct, competence or task performance. Today the traditional psychometric criteria of reliability and validity are further differentiated and extended with criteria of acceptability and utility (Messick, 1989; Moss, 1992; Stokking, Van der Schaaf, Jaspers & Erkens, 2004). Acceptability is about the assessment's objectivity, transparency, equality and unbiasedness, and utility is concerned with functionality, feasibility and efficiency.

In our view, for both summative and formative aims validity should have priority (Linn, 1994; Linn, Baker, & Dunbar, 1991; Borsboom, Mellenbergh, & Van Heerden, 2004), because assessment is primarily concerned with the intended competences. Transparency, unbiasedness, and functionality should also be satisfactory. In formative assessment, reliability, objectivity, and equality may receive less priority, because the results only indicate possibilities for improvement. Practical utility is also important, in order to be able to provide frequent and useful feedback. In summative assessment, reliability, objectivity and equality are also important (Stokking et al., 2004).

How portfolio assessments precisely can meet these criteria remains as yet unclear (Linn, 1994). In this study we focussed primarily on the reliability criterion, because reliability is a basic requirement and in portfolio assessment attaining acceptable levels of reliability has proved to be particularly difficult (Burns, 1999; Johnson, McDaniel & Willeke, 2000; Linn, 1994; Reckase, 1995; Shapley & Bush, 1999).

To estimate the interrater reliability of portfolio assessments different approaches are available, including different ways of summarizing ratings across raters, which may also impact the validity of the assessment results. Ideally the method of estimating interrater reliability is consistent with the underlying assumptions of the assessment study at hand. We shall now describe some available approaches and arguments for choosing among them.

Firstly, reliability is often estimated by the percentage of exact agreement in scoring between raters. Exact agreement is based on similarity of classification (Popping, 1983) and presumes that the classifications given are identical. This can only be realized when raters interpret the portfolio artefacts and the standards in completely the same way. As people always may differ in their interpretations (Huot, 1993; Kane, 1992; Van der Schaaf et al., 2003), these conditions cannot be guaranteed. Therefore, interrater reliability in terms of percentage agreement can only be used as an indicator of reliability.

Secondly, interrater reliability can be estimated by computing Cronbach's alpha coefficient for the raters' judgments on the criteria per portfolio (the so-called jury alpha). An underlying assumption is that two raters do not necessarily have to share a common interpretation of the standards and the portfolios, as long as each rater is consistent in classifying the portfolios according to his or her own interpretation of the standards. It is important to recognize that although the alpha may be high, the raters' judgment means may be different. Using correlations as an indicator for interrater reliability, we depart from the idea that reliable judgments should have similar mean scores *per se* (Murphy & De Shon, 2000). As subgroups of raters can have comparable interpretations, different interpretations within a group of raters are not necessarily idiosyncratic and rater variance effects do not necessarily indicate rater errors. That is, assuming the ratings are accurate. A condition for accuracy is that raters' judgments correspond with the performance theory (i.e., the standards) as taught in the training given to the raters (Borman, 1977).

Thirdly, reliability is also discussed using analysis of variance. Variability in scoring is generally based on, amongst other things, individual rater bias and systematic differences between raters, due to factors such as rater background and rater expectations. In generalizability studies, analysis of variance estimates the contribution of multiple sources of error to the score variance. Although some performance assessment studies report that the variance related to the judges makes a negligible contribution to the overall variance (e.g., Shavelson, Baxter, & Gao, 1993), this result does not occur when the assessment task is more complicated (Dunbar, Koretz, & Hoover, 1991).

As to the validity criterion of portfolio assessments, verbal protocols are often used to capture a rater's thoughts while assessing a teacher's portfolio. This is also relevant in our study. Two main forms of retrospective judgment protocol invalidity are reactivity and nonveridicality (Russo, Johnson, & Stephens, 1989). In the first case, a change of the primary cognitive processes occurs due to either the verbalization process or a prolonged response time. Nonveridicalities (inaccurate reflections on the underlying primary cognitive processes) contain errors of omission (not reporting some thoughts owing to overlooking or forgetting) and fabrication (reporting mental events that did not occur). Fabrication must be taken especially seriously because fabricated data are indistinguishable from other protocol data. Several researchers have warned against retrospective verbal protocols, and in particular stimulus-cued methods may lead to substantial fabrication (Ericsson & Simon, 1980, 1984; Russo et al., 1989). We assume that raters do not present a pure report of their underlying mental processes. Rather, they construct a representation of their judgments containing crucial information about the processes involved, including the communication within the context shared by the rater and the researcher (Long & Bourgh, 1996). In sum, verbal protocols provide useful and unique information about mental activities when fulfilling cognitive tasks, but they do not transparently reflect raters' mental processes when they are involved in silent assessment tasks. For that reason, verbal protocols are most valuable when used in combination with other data gathering instruments.

More details about the methods we used to further and control for the reliability and validity of the judgments in our study are described in the method section.

Raters' Cognitive Representations

Over the past decades, the assumption that raters' cognitive representations influence their judgments has been fed by social cognitive psychology. Several rating process models have been built that describe assessment activities (e.g., Landy et al., 1980; Feldman, 1981; DeNisi et al., 1984). Those models have in common that raters are using schemata while assessing, predicting and understanding another person's behaviour. These schemata are comparable to personal constructs (Kelly, 1955): content categories used to organize and simplify information. Interpersonal filters in these schemata induce raters to look selectively at the information concerning the people observed and to interpret this information according to their own constructs.

Rating process models as described in the literature are often related to Jones and Davis' (1965) Correspondent Inference Theory. This theory assumes that in the process of drawing inferences about others' behaviours, several activities occur (Gilbert, 1989). Raters categorize others' behaviour (e.g., "Irene speaks in a strict way") and relate this to feasible corresponding characteristics (e.g., "Irene must be a strict sort of person"). Meanwhile, they correct their characterizations by considering situational information (e.g., "Irene's students are very noisy, so Irene's strictness is reasonable. Maybe Irene is not such a strict person after all"). During the process, raters may express positive and negative utterances (Huot, 1993). Gilbert (1989) suggests that cognitive activities differ in their degree of consciousness. Our general impressions of others often remain implicit (Carlston, 1994). Further, the need to process information about the context of another's behaviour is not always obvious; when people are inattentive or unmotivated, they may fail to consider such situational information.

Consequently, when designing rating formats and training procedures it is important to be attuned to natural cognitive rating activities. Research into raters' judgment activities shows that it is possible to influence those activities. Various rater training procedures have been developed for this purpose (Lievens, 2001; Woehr & Huffcutt, 1994). Sceptics assume that rater training may only have minor effects on rater judgments because raters undergo socialization processes in their (professional) lives. Training periods and work experiences cannot easily replace mental representations formed during many years (cf. Huot, 1993).

Nevertheless, raters' cognitions form an essential domain that needs further exploration in order to understand better, and to improve, raters' judgmental processes (Day & Sulsky, 1995).

Associated Systems Theory

One way to study these judgmental processes is by using the Associated Systems Theory (AST) (Carlston, 1992, 1994). This theory focuses on several forms of human cognitive representations that operate simultaneously when forming impressions of other people. Furthermore, AST has proved to be usable for researching raters' judgments. For example, Schleicher and Day (1998) successfully used this theory to reveal impressions formed by assessors in an assessment centre setting.

The Associated System Theory is about person perceptions, impression formation, and social cognition. The theory focuses on the origin, organization and use of different kinds of mental representation of (other) people and events. Expanding on previous research in social psychology and neurology (Fiske, 1992; Martindale, 1991), two main beliefs within AST are *doing causes thinking* and *thinking is for doing*. The first principle says that people's cognitive representations develop by experiences derived from their actions. Translated to our study, raters' mental representations are influenced by their teaching experience, rater training, and experience in scoring portfolios. Secondly, mental representations can be seen as concepts that mediate the input of external stimuli and the output of actions (behavioural responses) (Norman, 1985). Thus, cognitive representations are fundamental to activities such as assessing portfolios.

AST offers a starting point for classifying assessors' representations. Carlston (1992, 1994) models AST on two dimensions (see Figure 1).

1. *Concrete versus abstract.* Concrete representations (left column) may include details of time and place of their recording (e.g., observing someone's physical appearance). The forms in the centre column recapitulate multiple observations and are therefore somewhat more abstract (e.g., attributing personal characteristics (e.g., lazy) based on someone's appearance). Abstract representational forms (right column) reflect general characteristics of the person and are not restricted to particular situations. Describing other persons in terms of personal characteristics in general is more cognitively demanding than providing concrete descriptions of someone's physical appearance. So this dimension also represents an expected increase of the cognitive activities needed.
2. *Target-referenced versus self-referenced.* Although mental representations are always more or less subjective and rater-bound, the strength of this varies. Target-referenced representations focus primarily on the target or person being assessed. Self-referenced representations concern rater's personal reactions to the assessee. Generally, personal reactions are based on a relatively stable cognitive structure and, compared to attitudes, may therefore be more difficult to change (Fazio, 1994). As raters always mentally interact with the assessment situation and with the assessee, assessing generally involves a mix of target-referenced and self-referenced representations (see middle row of Figure 1). Hence, their personal schemata filter their observations, their interpretions of their observations, and their final judgments (Tulving, 1983). So, even if an assessor does not physically interact with the situation and the person assessed, the assessor is always mentally involved.

	Concrete		Abstract
Target	(1a) Visual appearance	(1b) Categorizations	(1c) Attributed personality traits
	(2a) Behavioural observations	(2b) Mix	(2c) Evaluations
Self	(3a) Behavioural responses	(3b) Orientations	(3c) Affective responses

Figure 1: Structural Representation of the AST Taxonomy

Carlston (1992; 1994) specifies raters' cognitive representations as the cells in a matrix (Figure 1). Each cell represents a different type of cognitive representation.

1a Visual appearance: visual images of physical appearances are primary in our impressions of others. In our study, this is reflected in paraphrasing or referring to teachers' activities in the portfolio.

1b Categorizations (typing): sorting into sets and labelling of representations of others. In our study, this demands an interpretation or explanation of teachers' activities, as derived from the portfolio.

1c Attributed personality traits: describing others in terms of traits or character types. In our study, this refers to describing teachers' competences in accordance with the assessment criteria.

2a Behavioural observations: representations of others' behaviours are intermeshed with perceptions of their appearance and with features of one's own role in the recorded events. This category supposes that the rater mentally or physically interacts with the person assessed in a certain context, which is the case in assessments (Conway, 1990; Tulving, 1972, 1983).

2c Evaluations: verbal concepts used to embody affective feelings (e.g., "I think she has empathy with the students, which is good").

3a Behavioural responses: acts of the rater directed at the person assessed. This category is not relevant in our study because the raters do not interact with the teachers assessed.

3b Orientations: tendencies or predispositions to respond to a person in a particular manner; e.g., approach and avoidance tendencies of raters. This category is not researched in our study.

3c Affective response: "an abstract representation of affect that is linked to physiological structures involved in the primary experience of emotion" (Carlston, 1994, p. 6), e.g., laughing or crying. This category is not relevant in our study.

For validity reasons, it is important that raters alternately use concrete and abstract representations. One prerequisite for content validity is that concrete representations, such as observations of research assignments and video recordings, enhance the assessment's fit to the portfolio content. For the sake of acceptability of the judgment, in terms of its fairness and lack of ambiguity, the raters should also make clear on which concrete data the judgments were based. Furthermore, the assessment results should be suitable for the intended functions, such as providing comments on the teacher's strengths and weaknesses. Therefore, in the interest of practical utility, concrete examples that illustrate judgments are needed to give feedback to the teachers assessed. On the other hand, abstract representations are necessary for ratings to be accurate, according to the performance theory (the standards) as taught in the rater training. Abstract representations are also needed to predict with sufficient accuracy a teacher's performance in situations other than those described in the portfolio and in future employment. Furthermore, accurate interpretations of teachers' portfolios (interpretations that combine concrete and abstract representations) are needed to distinguish correctly between more and less competent teachers (the specificity aspect of validity) as well as to compare judgments on the same criterion based on different portfolio material (the convergence aspect of validity) (see Stokking et al., 2004).

We assume that an increase in the raters' use of target-referenced representations will increase the quality of their judgments of teachers' competences. The use of target-referenced representations increases the likelihood that the judgments are based primarily on teachers' competences and less on raters' subjective representations.

Standards for Teaching Students Research Skills

In one previous study we studied teachers' tasks needed to develop students' research skills and the competences (knowledge, skills and attitudes) needed to perform those tasks. We developed criteria that describe what teachers should know and be able to do in teaching students research skills (Van der Schaaf et al., 2004). Using a Delphi method, in three rounds 21 experts, including the raters in the present study, judged and revised a preliminary set of content standards. They judged the standards on a 4-point scale from *weak support* (score 1) to *strong support* (score 4). The method resulted in eight standards with a high degree of support. These criteria were:

1. Selecting goals for students (GOAL): teacher's long term goals for developing students' research skills and the approach s/he uses to reach these goals.
2. Choosing an appropriate assignment (ASSIGN): the goals, content and form of the assignment.
3. Preparing and managing students to work on their assignments (MANAGE): the preparation, communication, and management of facilities (time, rooms, sources, media, etc.) that students need for working on the assignment.
4. Planning and selection of teaching strategies that support the development of students' research skills (THINK): the choice of and argumentation for teaching strategies that meet students' knowledge, abilities and experience.
5. Teaching students research skills (TEACH): the use of the teaching strategies chosen.
6. Creating a positive pedagogical climate (CLIMATE): the provision of a safe, respectful, and stimulating learning environment for students.
7. Adequately assessing the research skills of students (ASSESS): teacher's argumentation for, and the clarity and comprehensibility of the assessment approach used (criteria, scoring, and norms), the applicability of the approach to the assignment, and the way the teacher handles the assessment and communicates the results to the students.
8. Reflecting on the program and on actions in teaching students research skills (REFLECT): the extent to which the teacher is aware of strong and weak points in his or her teaching, and his or her suggestions for improvement.

In a second previous study (Van der Schaaf et al., 2003), using policy capturing we developed the performance standards (the cut-off scores that indicate how good teachers should meet the content standards for a sufficient result). Policy capturing uses multiple regression procedures to develop the cut-off scores and to produce a best fitting equation or policy towards the weighting of the criteria and the specific judgment model. Nine experts, again including the raters in the present study, judged a great number of teacher profiles, which were simulated configurations of scores on the eight content standards. The judgments were analysed by linear multiple regression analysis, yielding a policy including the performance standards and the content standards' weights (the criteria TEACH and CLIM being given most weight).[1] This study resulted in three anchor points on the 5-point scales rating teachers' competences on the basis of their portfolios, namely that the teacher satisfies the criteria: *completely* (score 5), to a *reasonable degree* (score 3), or *not at all* (score 1). See Appendix 1 for an example.

Method

Sample of Raters

Having teaching experience themselves helps raters assess teachers' competences (Pula & Huot, 1993). We therefore selected raters with teaching experience. Preparing the selection process in the first-mentioned previous study (described above) we asked 20 stakeholders with different educational positions which kinds of persons should participate in the study. Most stakeholders mentioned practicing teachers and external experts such as teacher educators and pedagogical content experts. We chose to use external raters, that is raters not working in the same schools as the teachers to be assessed, in order to avoid raters biasing their perceptions or judgments on episodic memories and local orientations. We selected raters who successfully participated in the two previous studies (described above) in which content standards, performance standards, and a portfolio assessment procedure were developed (Van der Schaaf et al., 2003, 2004). So, the raters were familiar with the project and they subscribed to the standards and the assessment procedure (prerequisites for any assessment process to be succesful). The raters, who, as it happened, were all male, consisted of a school principal who was also an experienced teacher, two experienced geography teachers who were also teacher trainers, two experienced history teachers, an experienced teacher of both geography and history, and an experienced teacher of social science subjects who was also a teacher trainer in economics. The teacher trainers were working at several university teacher training institutes. All raters received financial compensation for participating. None of the raters had assessed teacher portfolios before participating in our project.

Sample of Teachers

From a random sample of 115 upper secondary schools in the Netherlands (20% of the population) we approached the heads of the economics, geography, and history departments with information about the study and a request for participation. Twenty-one teachers from twenty-one schools were willing to participate: 10 history, 7 economics, 4 geography. This low response can be explained by the demanding character of the study in terms of the required motivation, time, and experience in teaching students research skills.

Teachers' Portfolio Preparation

Teachers assembled the following artefacts during a period of a few months (cf. Figure 2).

1. Self-description, including a description of the teacher's experience and vision concerning teaching research skills.
2. A series of research assignments given to the students during the upper-secondary education years.
3. The results of two interviews concerning teachers' practical knowledge and intentions regarding instructing and coaching students research skills.
4. Two video recordings of lessons in which the teacher instructs and coaches students doing research.
5. Student evaluations of the teacher.
6. A research assignment that is central to the portfolio, including the learning goals of the assignment and the teacher's reasons for the assignment content and form.
7. The assessments of students' work by the teacher (including the goals, criteria, and scoring used).
8. Teachers' reflections on their weaknesses and strengths and on how to improve their teaching. For illustration purposes, the teachers also added examples of students' work.

We strengthened the generally weaker points of the portfolio instrument (low reliability and laboriousness) by specification of the portfolio content and context.

Criteria Portfolio Artefacts	GOAL	ASSIGN	MANAGE	THINK	TEACH	CLIMATE	ASSESS	REFL
1. Self-description	X							X
2. Series assignments	X							
3. Interviews			X	X	X	X		X
4. Videos			X		X	X		
5. Student evaluations		X	X		X	X		
6. Assignment	X	X	X		X		X	
7. Assessment							X	
8. Reflection								X

Figure 2: Criteria and Portfolio Artefacts

All teachers who participated in this study received an extensive report containing the judgments given (the scores to be described below) and also concrete feedback summarizing the contents of the completed judgment forms (see also below).

Rater Training Procedure

The raters studied a manual[2] that thoroughly described the objectives, planning and procedures of the assessment, the criteria, the anchor points, and the portfolio materials. The raters first individually studied an example portfolio. They then participated in a training session (four hours plenary) during which each rater independently practiced one analytic and one holistic assessment. The raters expressed their decisions on a judgment form on which they were encouraged to write comments. After individually scoring the portfolios, the group discussed their analytic and holistic assessments, with a focus on the arguments underlying the judgments given and their evidential base in the portfolio artefacts. The raters were taught how to use the rating format and to explain their arguments.

After the training, the raters individually rated three portfolios in a pilot (one randomly selected portfolio per subject), intended as a try-out of the rating procedure and to give the raters feedback in order to improve their judgments. In addition, we gathered judgment forms and retrospective verbal protocol data. We used these data to develop a coding scheme to analyse verbal protocols and judgment forms (see below). After the pilot, the raters received feedback about the reliability of their scoring and suggestions for improving their judgments. Suggestions included the accurate use of the standards, the assessment procedures, and the interpretation of the standards.

Instrumentation

We used three instruments to capture raters' cognitive representations: judgment forms, retrospective verbal protocols, and retrospective open-ended interviews.

Judgment Forms

Judgment forms were designed according to the Correspondent Inference Theory (Jones et al., 1965). Firstly, raters illustrated every criterion with significant portfolio material. Secondly, they described their interpretations of the material. Finally, raters assigned a score to each content standard, using a 5-point scale with anchor points (see Appendix 1) (hereafter: analytic scores). Raters added an overall judgment on a 5-point scale, considering the total of evidence from the separate scores on the content standards (hereafter: holistic score). Finally, for each portfolio, the raters reported whether they followed the prescribed procedure.

Retrospective Verbal Protocols

Raters explicated their thoughts in verbal protocols (Ericsson & Simon, 1984). Verbal protocols can be used concurrently (during the performance of a judgment task) or retrospectively (after the completion of the task). The first demonstrates the verbalization of raters' cognitive processes in short-term memory (thinking aloud) and the latter reveals mental representations that are stored in long-term memory (Ericsson & Simon, 1984). We used retrospective protocols by which raters recalled their thinking within two weeks after rating the portfolios, using their completed judgment forms as memory cues. We used retrospective protocols because portfolio assessment is time consuming (on average, four hours per portfolio) and mentally demanding, so concurrent verbalization would be too challenging.

Retrospective Open-Ended Interviews

Following the verbal protocol sessions, we conducted open interview sessions about raters' judgment approaches. The interviews focused on the procedure followed while judging the portfolios, the (cognitive) steps taken to build an overall impression of a teacher resulting in a holistic judgment score, pitfalls encountered during the assessment process, and the extent to which the judgments were target-referenced.

Coding Scheme

We used the pilot that was part of the training also to check for the quality of the scoring procedures and to develop a coding scheme to describe raters' cognitive representations. In the pilot, the six raters individually judged three portfolios (one per subject). In the judgment forms, all raters indicated that they followed the prescribed judgment procedure. The jury alphas of the analytic scores were .75, .44 and .63. In 50% of all holistic scores, the assessors showed exact total agreement, 33% of the scores showed a difference of 0.5 point, and 17% showed a difference of one or more points. These results are comparable to those obtained by LeMahieu, Gitomer, and Eresh (1995), who reported exact agreement among teachers' scores rating student portfolios ranging from 46% to 57%.

The eight criteria formed a consistent scale (Cronbach's alpha .79), which was a prerequisite for calculating unweighted and weighted mean analytic scores. A variance analysis did show slight rater variance. See Table 1.

Table 1: One-Way Analyses of Variance Between Teachers and Between Raters on Holistic Scores (H) and Weighted Mean Analytic Scores (Wa) (pilot, 3 portfolios)

Effects		*df*	Sum of squares	Mean square	*F*	(*p*)
Teachers	h	2	2.028	1.014	4.244	(.03)
	wa	2	4.819	2.409	19.146	(.00)
Raters	h	5	.944	.189	.486	(.78)
	wa	5	.432	.086	.165	(.97)

During the pilot, all six assessors retrospectively verbalized their judgments of one portfolio (from a history teacher) during a two-hour session with the researcher. Based on Ericsson et al. (1984), the raters were given general instructions to continue verbalizing while reconstructing their rating processes. The six protocols were taken at the raters' schools or private homes and were tape-recorded and fully transcribed. The raters were allowed breaks as needed. Despite the design of the judgment forms in accordance with the Correspondent Inference Theory, raters' judgment processes as retrospectively verbalized appeared to be associative, in that they skipped from one topic to another. We decided to separate segments per portfolio material per criterion (i.e., per cell in Figure 1). The intercoder agreement (Cohen's kappa) of the segmentation between two coders (the first author and an independent researcher) of three randomly chosen verbal protocols (three out of the six available) was .89. This result corresponds with the range of the reliabilities of segmentation between .80 and .90 as discussed in Ericsson et al. (1984).

The coding of the segments was primarily based on Associated Systems Theory. We used the codes *visual appearance*, *categorization* and *personal trait* to describe utterances on the dimension *concrete versus abstract*. We also accounted for the type of response given on the dimension *positive versus negative* (Huot, 1993). In accordance with the Correspondent Inference Theory, as a third category we encoded verbalizations about the teacher's situation. To check this initial coding scheme, the two coders coded the three verbal protocols from the pilot (the same two coders and three randomly chosen protocols as mentioned above). Discussions between the two coders resulted in some adjustments to the initial scheme. To accurately capture raters' individual representations on the dimensions *concrete-abstract* and *positive-negative* we decided to encode the protocols per line. In addition, we added two more categories: *rater's own judgment process* and *judgment procedure*. Finally, to facilitate reliable coding, we developed discourse markers for every category. In sum, we coded on the dimensions *concrete-abstract* and *positive-negative* per line, and on three other categories, *situation in which teachers operate*, *raters' own judgment process*, and *judgment procedure* per segment (see Appendix 2).

Table 2: Cohen's Kappas of Categorizations Based on the Retrospective Verbal Protocols and Judgment Forms

Category of comments	Verbal protocols	Judgment forms
Dimension abstract–concrete (encoding per line)	.71	.65
Dimension positive–negative (encoding per line)	.66	.79
Situation in which teachers operate [a]	.75	1.0
Raters' own judgment process (metacognitive) [b]	.69	.48
Judgment procedure [c]	.65	.80

Notes:

[a] Subcategories: own background, method used, students' characteristics, time available, teacher's experience, school policy, expectations.

[b] Subcategories: references to the manual while rating; notifications about lack of information for giving an adequate assessment; explaining own judgment process; pitfalls and uncertainties while rating; questioning own assessment process; notifications and corrections of mistakes; references to earlier assessed portfolio material; references to earlier assessed portfolios.

[c] Subcategories: criteria and standards, portfolio material, assessment procedure, weighting of the criteria.

The five categories included in the final coding system could satisfactorily be distinguished in the verbal protocols and the judgment forms (see Table 2 above). With one exception, the kappas were above .65, which is acceptable for intercoder agreement (Popping, 1983).

To obtain scores on the dimensions *abstract–concrete* (*visual appearance, categorization, attributed personality traits,* coded as 1, 2, 3) and *positive–negative* (*positive, neutral, negative,* also coded as 1, 2, 3) we averaged the rater's positions on both dimensions and then averaged the scores across the protocol lines per portfolio per rater. The frequencies on the other three categories (*teachers' situations, raters' judgment processes,* and *judgment procedures*) were counted per portfolio per rater per subcategory.

Data Collection

After having used three portfolios in the pilot as described above, the remaining 18 portfolios (9 history, 6 economics, 3 geography) were rated, each one independently by two raters. We organized the rater pairs according to their major expertise. As most of the raters had expertise in several subjects, the composition of the rater pairs could vary according to availability of the raters. In a 9-month period, we assigned the 18 portfolios in the order they became available. Rater 1 judged six portfolios (mainly economics), Rater 2 four portfolios (history), Rater 3 nine portfolios (mainly history), Rater 4 five portfolios (geography and history), Rater 5 seven portfolios (geography, history and economics), and Rater 6 five portfolios (mainly economics). We aimed at using fixed couples per subject, but due to some availability problems we had to allow for a few departures from this scheme.

All raters retrospectively verbalized their rating process of two randomly chosen portfolios (after scoring) from the portfolios they had judged. The protocols were fully transcribed, resulting in 7216 lines containing 310 segments (split per content standard per portfolio, according to the cells in Figure 2). The independent coder (the one who earlier showed intercoder agreement with the researcher) coded the verbal protocols of 12 portfolios (two portfolios per rater) and the 36 judgment forms of all 18 portfolios, using the coding scheme developed earlier.

Data Analysis

We describe the analyses concerning the reliability of raters' judgments, raters' cognitive representations, and the relationships among raters' cognitive representations, their judgments, and the reliability of these judgments according to the three research questions.

The Reliability of the Raters' Judgments

The rating produced three types of score: unweighted mean analytic, weighted mean analytic, and holistic. The unweighted mean analytic scores focused on the mean of the scores on the eight content standards. The weighted analytic scores used the weights per content standard to calculate total scores (see Note 1). The holistic scores were overall judgments considering the total of evidence from the separate scores on the content standards.

Firstly, we analysed the reliability of the eight content standards as a set by computing Cronbach's alpha. Secondly, we analysed the extent of agreement in holistic judgments within rater pairs. As an additional indication of the consistency of raters' judgments, we verified whether there were statistically significant differences between the means of the three score types: unweighted mean analytic, weighted mean analytic, and holistic.

Thirdly, we examined the interrater reliability for the separate analytic scores by calculating the Cronbach's alphas (jury alphas). The accuracy of the ratings, that is the correspondence between raters' judgments and the performance theory as taught in the training, was checked by coding the content of raters' judgments and comparing the results with the full descriptions of the standards in our study.

Finally, we conducted variance analyses on the unweighted and the weighted mean analytic scores and the holistic scores to estimate the variance related to the raters.

Raters' Cognitive Representations

Data from the retrospective interviews were qualitatively analysed. The coded data from the verbal protocols and judgment forms were analysed in a descriptive way (frequencies and means) and the differences between raters were analysed using one-way analysis of variance.

Relationships Between Raters' Cognitive Representations, their Judgments and the Reliability of these Judgments

The assumption that judgments are a function of cognitive representations entails the idea that raters who differ in their comments should also differ in their judgments. We used linear multiple regression analysis (method enter) to explore how well the scores on the categories of the comments given in the judgment forms and the retrospective verbal protocols predicted the unweighted mean analytic, weighted mean analytic, and holistic scores. In the analyses of the judgment forms we eliminated the category *rater's own judgment process*, because this category was not satisfactorily distinguished during the coding process (see Table 2). To indicate the quality of the prediction, we used the proportion of explained variance (R^2). For linear multiple regression analyses to be valid, assumptions of normality of residuals, homoscedasticity, linearity, and absence of collineation have to be fulfilled. Therefore we checked for these assumptions.

Finally, to check whether the raters' cognitive representations possibly influenced the reliability of their portfolio judgments, we calculated the Pearson correlations between the differences in the category scores within rater pairs and the jury alphas.

Results

The Reliability of the Raters' Judgments

The eight criteria formed a reliable scale: the Cronbach's alpha was .76. To 35% of the judgments the raters gave exactly the same holistic score on the 5-point scale, a difference of half a point showed up in 12% of the judgments, a difference of one point in 47% of the judgments, and a difference of 1.5 points in 6% of the judgments (one rater pair).

The jury-alphas for the separate analytic scores were satisfactory for 12 rater pairs, ranging from .39 to .76. The jury-alphas were low or even negative for six pairs, ranging from -.80 for one pair and ranging from -11 to .22 for five pairs. We compared the content of raters' judgments forms to the full descriptions of the content standards. In all ratings, the raters referred sufficiently to the content standards according to the training.

To check the agreement between the three types of scores, we calculated correlations and conducted paired-samples *T*-tests (see Table 3). The results showed that the score types were highly correlated. However, the mean of the holistic scores (3.12; *SD* .78) is significant lower than the mean of the weighted mean analytic scores (3.60; *SD*. .96), while the former corresponds closely with the mean of the unweighted mean analytic scores (3.11; *SD* .59).

To estimate whether the scoring variability was due to rater effects, we conducted one-way analyses of variance (see Table 4). The results showed slight rater effects. Due to the structure of the available data we could not check for possible interaction effects between raters and teachers.

Table 3: Pearson Correlations and Paired Samples *T*-tests for Holistic (h), Weighted Mean Analytic (wa), and Unweighted Mean Analytic (ma) Scores (18 portfolios)

	First mean	Second mean	r^2	Difference (*sd*)	*df*	*T*	(*p*)
h – wa	3.12 (h)	3.60 (wa)	.85 *	–.44 (.50)	35	–5.28	(.00)
h – ma	3.12 (h)	3.11 (ma)	.85 *	.04 (.42)	35	0.59	(.56)
wa – ma	3.60 (wa)	3.11 (ma)	.99 *	.49 (.38)	35	7.59	(.00)

* p<.00

Table 4: One-Way Analyses of Variance Between Teachers and Between Raters on Holistic (h), Weighted Mean Analytic Scores (wa), and Unweighted Mean Analytic (ma) Scores (18 portfolios)

Effects		*df*	Sum of squares	Mean square	*F*	(*p*)
Teachers	h	17	12.90	.76	1.54	(.17)
	wa	17	19.64	1.16	1.67	(.14)
	ma	17	7.40	.44	1.62	(.16)
Raters	h	5	.97	.19	.29	(.92)
	wa	5	2.04	.41	.41	(.84)
	ma	5	1.07	.21	.58	(.72)

Raters' Cognitive Representations

Retrospective Interviews

In the retrospective interviews, the raters said they used the rating procedure according to their training. They also said they found the procedure useful. They spent an average of four hours per portfolio to produce all judgments and ratings.

Five raters (1, 3, 4, 5, 6) primarily strove to build a coherent image of the teacher based on the portfolio material. They subsequently explained this image using concrete examples. In building an image, they perceived the videos as very helpful. Four raters firstly watched the videos and made notes (filtering out relevant data). Then they went through the portfolio. Finally, they filled out the judgment forms and went through the portfolio per content standard for a second time. In an alternative process, two raters directly started filling out the judgment form. They went through the videos and portfolios only once. *T*-tests between these two raters and the other raters showed that these variations in procedure did not significantly affect the rating outcomes.

Some raters discovered that previously rated portfolio material and content standards influenced the rating process of later portfolio material and content standards. Rater 1 was aware of this process: "I should not use the self-description to assess Criterion 3 [MANAGE], but sometimes it stays in my head and it influences my rating, although it should not". Most of the raters said that to control for this, they corrected this process afterwards. Rater 4: "In rating teachers, I have a tendency to follow my intuition. Since I know that is not correct, I force myself to follow the procedure as trained conscientiously". Rater 5 explained the need to be aware of the danger of looking for episodes in the portfolio that confirmed his preconceptions. "Every time I correct myself for this mechanism by saying to myself, 'Watch out, you are doing it again'". Rater 2: "A few days after assessing a portfolio, I re-read my assessment to control if it really justifies the portfolio".

The images built were also influenced by the raters' own experiences. Four raters (1, 3, 5, 6) confirmed that they based their judgments partly on their own knowledge about teachers. Rater 3: "You know what is realistic to expect from teachers. Carrying your experiences with you colours the judgments given".

Judgment Forms and Retrospective Verbal Protocols

Table 5 shows that all raters used a mix of concrete and abstract representations in their comments. On average, the verbal protocols elicited more comments from the raters than the judgment forms, especially on the dimension *concrete–abstract*. In general, comments made in the verbal protocols were less concrete than comments in the judgment forms. Furthermore, the raters differed in the number of comments they made per portfolio.

One-way analyses of variance of raters' mean scores on the dimensions *concrete–abstract* and *positive–negative* in judgment forms and verbal protocols showed that the raters differed in their mean scores. The results on the dimension concrete–abstract were $F = 4.19, p < .001$ (judgment forms), $F = 0.06, p < .001$ (verbal protocols). On the dimension positive–negative, the results were $F = 6.24, p < .001$ (judgment forms), $F = 0.04, p < .001$ (verbal protocols). Other differences in their comments were not significant.

Table 5: Means and Frequencies of Comments per Rater in Verbal Protocols and Judgment Forms

Sources		Concrete Abstract			Positive Negative			Situation teacher	Own judgment	Judgment procedure
Rater	*n*	*m*	*sd*	*n*	*m*	*sd*	*n*	*n*	*n*	*n*
1										
Verbal protocol	2	1.74	.70	530	2.69	.67	62	42	14	108
Judgment form	6	1.54	.57	357	1.98	.98	56	10	49	7
2										
Verbal protocol	2	1.96	.60	138	2.31	.75	20	1	0	13
Judgment form	4	1.51	.54	219	2.13	.95	24	6	30	4
3										
Verbal protocol	2	1.89	.66	1079	2.34	.84	128	50	29	104
Judgment form	9	1.44	.53	1456	1.77	.94	176	17	72	19
4										
Verbal protocol	2	1.91	.68	554	1.86	.89	94	26	33	80
Judgment form	5	1.55	.56	302	2.32	.88	78	6	11	7
5										
Verbal protocol	2	1.95	.60	594	1.97	.90	68	37	27	87
Judgment form	7	1.45	.51	504	1.74	.97	76	14	26	2
6										
Verbal protocol	2	1.94	.64	613	2.47	.80	83	31	23	79
Judgment form	5	1.47	.53	852	2.16	.91	156	9	22	1

Relationships between Raters' Cognitive Representations, their Judgments and the Reliability of these Judgments

To explore whether the categories of comments could explain the judgments given, we conducted linear multiple regression analysis, using raters' unweighted mean analytic, weighted mean analytic, and holistic scores as the criterion variables and the scores on the categories of cognitive representations as predictors. For all three score types the assumptions of normality, homoscedasticity and linearity were sufficiently met. The categories of cognitive representations based on the retrospective verbal protocols ($n = 12$, concerning two randomly chosen portfolios per rater) did not show significant relations with the judgment scores, but those based on the judgment forms (n = 36) did. The explained variances of the unweighted mean analytic scores (R^2 = .65, F = 2.59, df = 13, p = .03), the weighted mean analytic scores (R^2 = .63, F = 2.68, df = 13, p = .03), and the holistic scores (R^2 = .59, F = 2.36, df = 13, p = .04) were quite similar.

Linear multiple regression analysis of the unweighted mean analytic scores on the categories of comments in the judgment forms (n = 36) showed that 65% of the variance in the ratings could be explained by the category scores. The category scores were barely collineated (the correlations were between – .22 and +.31 and mostly near r = .00). The standardized partial beta coefficients showed that the comments made on the dimensions *concrete–abstract* (beta = .35; t = 2.13; p = .05) and *positive–negative* (beta = .71; t = 4.4; p = .00) contributed significantly to the prediction of the portfolio judgments given. The other categories (concerning the teacher's situation and the judgment procedure) did not make a statistically significant contribution to the prediction.

Since the scores on the cognitive representation dimensions *concrete–abstract* and *positive–negative* significantly contributed to the prediction of the judgments given, we also checked their relationship with the interrater reliability of these judgments. Per portfolio (n = 18), we calculated the difference scores between the two raters on both dimensions and then we correlated these difference scores with the jury alphas. The results showed that the closer the correspondence within the rater pairs' comments in the judgment forms as categorized on the two dimensions, the higher their interrater reliability (r = .44, p = .09 for the dimension *concrete–abstract*; r = .53, p = .04 for the dimension *positive–negative*).

Conclusion and Discussion

We researched into the questions: What is the reliability of teacher portfolio assessments? How can raters' cognitive representations be described? What is the relationship between raters' cognitve representations, their judgments and the reliability of these judgments? We used a mixed quantitative and qualitative approach to study the reliability of judgments, raters' cognitive representations (using the Correspondent Inference Theory (Jones & Davis, 1965) and the Associated Systems Theory (Carlston, 1992, 1994)), and the relationships between cognitive representations, judgments and reliability.

Six raters systematically assessed 18 portfolios. The interrater reliability of the analytic scores of 12 portfolios was satisfactory. Variance analyses showed only slight rater effects. Raters' holistic scores in all but one case differed by no more than one point (on a 5-point scale).

The raters used cognitive representations on the dimensions *concrete-abstract* (e.g., visual manifestations) and *positive–negative* (positive and negative comments), the situation in which teachers operate, their own judgment process, and the judgment procedure.

Although we used the judgment forms completed during the judgment sessions as cues for producing the retrospective verbal protocols, the comments in the verbal protocols differed markedly from those in the judgment forms. The comments made in the verbal protocols were generally less concrete than those in the judgment forms. The cognitive representations in the judgment forms were significantly related to the judgments given. The differences in cognitive representations within pairs were significantly related to the interrater reliabilities.

What is the meaning of these results? Firstly, several researchers have argued that the complex and context-bound character of portfolio assessments asks for the use of a different reliability standard than used for standardized forms of assessment in which interrater reliability coefficients often exceed .90 (Nunnally, 1978). After all, many variables that are fixed in standardized tests are variable in portfolio assessments. Koretz, Klein, McCaffrey, and Stecher (1992) assert that interrater reliability coefficients of .80 or higher are often regarded as reasonably strong for performance assessments. Gentile (1992) reports that coefficients above .80 strong, and above .65 are considered good for portfolio studies. We could only partly meet these criteria. Further research should expose the coefficients attainable for portfolio assessments.

Secondly, the raters based their holistic judgments of the portfolios on the unweighted means of the analytic scores given. This is obvious from the correspondence between these two score types. This tendency could be considered as undesirable, however, since it is possible that one criterion is more important than another (and thus should get more weight). In one of our two previous studies (Van der Schaaf et al., 2003) we let the raters formulate a panel policy in which they decided that some criteria should be weighted more heavily than others. The tendency to use unweighted means is strong, however, as the holistic scores and the unweighted mean analytic scores corresponded highly although the raters were instructed and trained in weighting the scores according to the policy.

Thirdly, the retrospective interviews revealed that the raters strove to build a coherent image of the teacher they were assessing. The raters' experience was that the images were influenced by previously rated portfolio material and content standards, as well as raters' own experiences as a teacher. These results are comparable with those of other studies in diverse domains. Zwaan and Brown (1996), for example, found that skilled readers generate explanations to combine data across sentences, building a coherent mental representation.

Fourthly, a prerequisite for valid assessments is that raters base their judgments on teachers' portfolio content. Rating quality is improved by raters using target-referenced representations, relating judgments primarily to teachers' competences rather than to raters' subjective, self-referenced perceptions. Forming concrete representations enhances this process. On the other hand, abstract representations are needed to predict a teacher's performance in situations other than those described in the portfolio. Ideally then, raters should alternately use concrete and abstract representations.

Fifthly, the judgments given could significantly be predicted by the raters' representations as categorized on the basis of their comments in the judgment forms on the dimensions concrete–abstract and positive–negative. These representations explained a high 65% of the variance in the portfolio ratings. Furthermore, the results showed that the higher the correspondence within rater pairs in these representations, the higher the interrater reliability. These dimensions of raters' cognitive representations as revealed in our study could therefore also explain the reliability of the assessment.

Since portfolios are personal and context-bound instruments, portfolio assessment is cognitively demanding and complex, even after training. This affects the reliability of portfolio assessments (expressed in the interrater reliability and rater effects). Although our study showed that certain cognitive representations may contribute to reliable portfolio assessment, it also shows that raters differ in their representations, which seemingly reflects different interpretations of the portfolios, regardless of the agreement in the ratings given.

These results give way to the question whether consensus approaches would be preferable as an assessment strategy and whether such approaches could replace independent ratings. The value of consensus approaches depends on the importance of interrater consistency in portfolio assessment. For summative assessment purposes—e.g., certification or merit pay—portfolios contain a sample of "best work" and high interrater reliability will be necessary to make adequate decisions. However, for formative purposes and using portfolios that contain broad and various samples of work, it can be questioned whether it is likely that two independent raters will arrive at identical judgments. In this case, it might be more important that raters agree on the consequences of the judgment given, rather than on the actual scores achieved.

Delandshere and Petrosky (1994) proposed a procedure using confirmation of prior judgments (similar to getting a second opinion from a physician) rather than replications. That such a confirmation approach might be useful for portfolio assessment was also concluded by Linn (1994), refering to the work of Moss (1994) on validity versus reliability and illustrating how committees examining the qualifications of candidates can arrive at an integrated decision. Further research should demonstrate the consequences and practical utility of confirmation approaches.

Portfolio assessments are contextually embedded. Our study revealed that external raters involve situational information (e.g., student characteristics and school policy) in their rating processes. They may partly color this information with their own teaching experiences. Internal raters, working in the same educational organization as the teachers being assessed, could ensure more validity and consistency in their ratings. It is often suggested that individuals from similar backgrounds are likely to reach agreement more readily (Pula & Huot, 1993). For that reason, portfolio assessment may be more suitable for internal rating processes (Linn, 1994). However, internal ratings of teacher portfolios will probably echo subjective impressions and personal relationships between the rater and the teacher assessed. As a result, the ratings will be more idiosyncratic and this will have its own negative effects on the rating reliability. It can be expected that internal ratings of teacher portfolios will be more self-referenced than those of external raters (Carlston, 1994). To increase the accuracy of internal ratings, rater training should pay attention to the importance of target-referenced rating. Future research should examine such possibilities.

Acknowledgement

This study was funded by the Dutch Organisation for Scientific Research (NWO-PROO 490-23-081)

Notes

1. The linear multiple regression analysis in the second previous standard setting study resulted in the following weights (standardized partial betas) per content standard: GOAL 0.16, ASSIGN 0.23, MANAGE 0.06, THINK 0.15, TEACH 0.33, CLIMATE 0.33, ASSESS 0.19, REFL 0.14.
2. The manual is available upon request.

References

Andrews, T.E., & Barnes, S. (1990). Assessment of teaching. In W.R. Houston (Ed.), *Handbook of research on teacher education* (pp. 569-598). New York: Macmillan.

Bird, T. (1990). The schoolteacher's portfolio. In J. Millman & L. Darling-Hammond (Eds.), *The new handbook of teacher evaluation: Assessing elementary and secondary school teachers* (pp. 241-256). Newbury Park, CA: Sage.

Borman, W.C. (1977). Consistency of rating accuracy and rating error in the judgment of human performance. *Organizational Behavior and Human Performance, 20,* 238-252.

Borsboom, D., Mellenbergh, G.J., Van Heerden, J. (2004). The concept of validity. *Psychological Review, 111* (4), 1061-1071.

Boud, D. (1990). Assessment and the promotion of academic values. *Studies in Higher Education, 15,* 101-111.

Burns, C.W. (1999). Teaching portfolios and the evaluation of teaching in higher education: Confident claims, questionable research support. *Studies in Educational Evaluation, 25,* 131-142.

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* (pp. 301-341). Hillsdale, NJ: Erlbaum.

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

Cochran-Smith, M., & Fries, M.K. (2001). Sticks, stones, and ideology: The discourse of reform in teacher education. *Educational Researcher, 30,* 3-15.

Conway, M.A. (1990). Associations between autobiographical memories and concepts. *Journal of Experimental Psychology: Learning, Memory and Cognition, 16,* 799-812.

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.

De Corte, E. (1996). Instructional psychology: Overview. In E. De Corte & F.E. Weinert (Eds.), *International encyclopedia of developmental and instructional psychology* (pp. 33-43). Oxford: Elsevier Science.

Delandshere, G., & Arens, S.A. (2001). Representations of teaching and standards-based reform: are we closing the debate about teacher education? *Teaching and Teacher Education, 17,* 57-566.

Delandshere, G., & Petrosky, A. (1994). Capturing teachers' knowledge: Performance assessment (a) and post-structuralist epistemology, (b) from a post-structuralist perspective, (c) and post-structuralism, (d) none of the above. *Educational Researcher, 23,* 11-18.

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.

Dierick, S., & Dochy, F. (2001). New lines in edumetrics: New forms of assessment lead to new assessment criteria. *Studies in Educational Evaluation, 27,* 307-329.

Dunbar, S.B., Koretz, D., & Hoover, H.D. (1991). Quality control in the development and use of performance assessments. *Applied Measurement in Education, 4,* 289-304.

Ericsson, K.A., & Simon, H.A. (1980). Verbal reports as data. *Psychological review, 87,* 215-251.

Ericsson, K.A., & Simon, H.A. (1984). *Protocol analysis: Verbal reports as data.* Cambridge, MA: MIT Press.

Fazio, R.H. (1994). Attitudes in associated systems theory. In R.S. Wyer (Ed.), *Associated systems theory: A systematic approach to cognitive representations of persons. Advances in social cognition* (pp. 157-167). Hillsdale, NJ: Erlbaum.

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

Fiske, S.T. (1992). Thinking is for doing: portraits of social cognition from daguerreotype to laserphoto. *Journal of Personality and Social Psychology, 63,* 877-889.

Gentile, C. (1992). *Exploring new methods for collecting students' school-based writing: NAEP's 1990 Portfolio Study.* Washington, DC: Office of Educational Research and Improvement.

Gilbert, D.T. (1989). Thinking lightly about others: Automatic components of the social inference process. In J.S. Uleman & J.A. Bargh (Eds.), *Unintended thought* (pp. 189-211). New York: Guilford.

Huot, B.A. (1993). The influence of holistic scoring procedures on reading and rating student essays. In M.M. Williamson & B.A. Huot (Eds.), *Validating holistic scoring for writing assessment. Theoretical and empirical foundations* (pp. 206-232). Cresskill, NJ: Hampton Press.

Jones, E.E., & Davis, K.E. (1965). From acts to dispositions: the attribution process in person perception. In L. Berkowitz (Ed.), *Advances in Experimental Social Psychology, 2* (pp. 219-266). New York: Academic Press.

Johnson, R.L., McDaniel, F., & Willeke, M.J. (2000). Using portfolios in program evaluation: An investigation of interrater reliability. *The American Journal of Evaluation, 21,* 65-80.

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

Kelly, G.A. (1955). *The psychology of personal constructs.* New York: Norton.

Koretz, D., Klein, S., McCaffrey, D., & Stecher, B. (1992). *The reliability of scores from the 1992 Vermont portfolio assessment program.* Washington, DC: RAND Institute on Education & Teaching.

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

LeMahieu, P., Gitomer, D., & Eresh, J. (1995). Portfolios in large-scale assessment: Difficult but not impossible. *Educational Measurement: Issues and Practice, 14,* 11-16, 25-28.

Lievens, F. (2001). Assessor training strategies and their effects on accuracy, interrater reliability, and discriminant validity. *Journal of Applied Psychology, 86,* 255-264.

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

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

Long, D.L., & Bourgh, T. (1996). Thinking aloud: telling a story about a story. Commentary. *Discourse Processes, 21,* 329-339.

Lyons, N. (Ed.) (1998). *With portfolio in hand. Validating the new teacher professionalism.* New York: Teachers College Press.

Martindale, C. (1991). *Cognitive psychology: A neural-network approach.* Pacific Grove, CA: Brooks/Cole.

Messick, S. (1989). Validity. In R.L. Linn (Ed.). *Educational measurement* (pp. 13-103). New York: MacMillan.

Moss, P.A. (1992). Shifting conceptions of validity in educational measurement: Implications for performance assessment. *Review of Educational Research, 62* (3), 229-258.

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

Murphy, K.R., & De Shon, R. (2000). Interrater correlations do not estimate the reliability of job performance ratings. *Personnel Psychology, 53,* 873-900.

Norman, D.A. (1985). Human information processing: the conventional view. In A.M. Aitkenhead & J.M. Slack (Eds.). *Issues in cognitive modelling* (pp. 309-336). Hillsdale, NJ: Erlbaum.

Nunally, J. (1978). *Psychometric theory* (2nd ed.). New York: McGraw-Hill.

Pula, J.J., & Huot, B.A. (1993). A model of background influences on holistic raters. In M.M. Williamson & B.A. Huot (Eds.), *Validating holistic scoring for writing assessment. Theoretical and empirical foundations* (pp. 237-265). Cresshill, NJ: Hampton Press.

Popping, R. (1983). *Overeenstemmingsmaten voor nominale data* [Computing agreement on nominal data}. (Unpublished doctoral dissertation), Groningen, the Netherlands: Rijksuniversiteit Groningen.

Reckase, M.D. (1995). Portfolio assessment: a theoretical estimate of score reliability. *Educational Measurement: Issues and Practice, 14*, 12-14, 31.

Russo, J.E., Johnson, E.J., & Stephens, D.L. (1989). The validity of verbal protocols. *Memory & Cognition, 17*, 759-769.

Schleicher, D.J., & Day, D.V. (1998). A cognitive evaluation of frame-of-reference rater training: Content and process issues. *Organizational Behavior and Human Decision Processes, 73*, 76-101.

Shapley, K.S., & Bush, M.J. (1999). Developing a valid and reliable portfolio assessment in the primary grades: building on practical experience. *Applied Measurement in Education, 12*, 111-132.

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

Stokking, K., Van der Schaaf, M., Jaspers, J., & Erkens, G. (2004). Teachers' assessment of students' research skills. *British Journal of Educational Research, 30*, 93-115.

Tulving, E. (1972). Episodic and semantic memory. In E. Tulving & W. Donaldson (Eds.), *Organization and memory* (pp. 381-403). New York: Academic Press.

Tulving, E. (1983). *Elements of episodic memory*. Oxford: Clarendon Press.

Van der Schaaf, M.F., Stokking, K.M., & Verloop, N. (2003). Developing performance standards for teacher assessment by policy capturing. *Assessment & Evaluation in Higher Education, 28*, 395-410.

Van der Schaaf, M.F., Stokking, K.M., & Verloop, N. (submitted). Developing teaching content standards using a delphi method.

Woehr, D.J., & Huffcutt, A.I. (1994). Rater training for performance appraisal: A quantitative review. *Journal of Occupational and Organizational Psychology, 67*, 189-205.

Zwaan, R.A., & Brown, C.M. (1996). The influence of language proficiency and comprehension skill on situation-model construction. *Discourse Processes, 21*, 289-327.

The Authors

MARIEKE VAN DER SCHAAF is an educational researcher and a teacher in Educational Sciences at the Department of Educational Sciences, Utrecht University, the Netherlands. Her main areas of research include teacher evaluation and teaching and assessment of students' general and academic skills.

KAREL STOKKING is a professor of education at the Department of Educational Sciences, Utrecht University, the Netherlands. His areas of interest include the development and assessment of general and academic skills in secondary and higher education.

NICO VERLOOP is professor in education and director of ICLON Graduate School of Education, Leiden University, the Netherlands. He is immediate past president of the Dutch Educational Research Association. His major research interests are: teachers' practical knowledge, learning and professional development of teachers and teacher evaluation.

Correspondence: <m.f.vanderschaaf@fss.uu.nl>

Appendix 1

Content standard description: An example

Content standard 5 (TEACH): Teaching students research skills

Score 5, the teacher satisfies the criterion completely: the teacher uses multiple teaching strategies that support the development of student understanding in conducting investigations, and challenges students to accept and share responsibility for their own learning (see Standard 4).

He or she uses the strategies in an adequate manner. For example, if students collaborate, he or she recognizes and responds to students' diversity and encourage all students to participate fully in fulfilling the assignment. He or she encourages each student to share his or her ideas with the other members.

The strategies he or she uses suit the phases of the investigations the students are working at. For example, in the phase of formulating a research question, he or she coaches students in brainstorming; in the phase of gathering information, he or she gives the students keywords for searching the internet; in the phase of data processing, he or she directs the students to work in groups.

The teacher monitors students' understanding and the amount of direction given in the teaching/learning process. In order to gather data, he or she can observe students, listen to students, read their logs. He or she uses the data to improve his or her teaching practice.

Score 4, the teacher satisfies the criterion to a high degree.

Score 3, the teacher satisfies the criterion to a reasonable degree: the teacher uses a (few) teaching strategies that support the development of student understanding in conducting investigations, or challenges students to accept and share responsibility for their own learning. He or she uses the strategies in a reasonable manner.

The strategies he or she uses may suit the phases of the investigations the students are working at. The teacher sometimes monitors students' understanding and the amount of direction given in the teaching/learning process.

Score 2, the teacher satisfies the criterion to a small degree.

Score 1, the teacher does not satisfy the criterion: the teacher does not use teaching strategies that support the development of student understanding in conducting investigations, and does not challenge students to accept and share responsibility for their own learning. The strategies he or she uses do not suit the phases of the investigations the students are working at. The teacher does not monitor students' understanding and the amount of direction given in the teaching/learning process.

Appendix 2

Categories of comments

Dimension concrete-abstract (encoding per line)

1. Visual appearance (e.g., "She says ...")
2. Categorizations (e.g., "That means ...")
3. Personality traits (e.g., "She is an organized person.")

Dimension type of comment (encoding per line)

1. Negative (all sorts of negative comments)
2. Neutral (neutral or balanced feedback; e.g., "On the one hand [followed by a negative comment]; on the other hand [followed by a positive comment]")
3. Positive (all sorts of positive comments)
4. Tips and suggestions (e.g., "He rather should ...")

Situation in which teachers operate (encoding per segment)

1. To place oneself in the teacher's situation proceeding from his/her own experiences and background*
2. Comments regarding the material and method used by the teacher. Taken into consideration while rating:
3. Students' characteristics
4. Time available to the teacher
5. Teacher's experience in teaching students research skills
6. Teacher's collaboration with colleagues*
7. The school policy
8. General expectations towards teachers
9. The possibility of bias due to the researcher's presence in the school context*

Raters' own judgment process (encoding per segment)

1. References to the manual while rating (e.g., "the manual says ...")
2. Notifications about lack of information for giving an adequate assessment (e.g., "She says she gives feedback to her students, but in her portfolio I do not find any information about that.")
3. Explaining own judgment process (e.g., "I am rating in this way, because ...")
4. Pitfalls and uncertainties while rating (e.g., "The difficulty is ...")
5. Questioning own assessment process (to the researcher) (e.g., "Am I doing it right now?")*
6. Notifications and corrections of mistakes (e.g., "What I am doing now is not correct.")
7. References to earlier assessed portfolio material
8. References to earlier assessed portfolio(s)
9. Other metacognitive comments

Judgment procedure (encoding per segment)

Remarks (positive and negative) about:
1. Judgment criteria and standards
2. Portfolio material
3. Assessment procedure
4. Weighting of the criteria
5. Other

* This category does not occur in the judgment forms.