Imaging and assessing teacher competence: design characteristics of structured video portfolios

Erik Roelofs (Cito, Arnhem, The Netherlands)
Mirjam Bakker (Leiden University, The Netherlands)
Ellen Van Den Berg (University of Twente, The Netherlands)

Paper presented at the annual meeting of the American Educational Research Association, April 17th San Diego (CA).

INCOMPLETE DRAFT VERSION, FOR DISCUSSION PURPOSES ONLY

1. Introduction

In teacher preparation programs an increasing emphasis is laid on (authentic) performance assessments (Darling-Hammond, 2000). In these assessments learner teachers perform realistic and complex teaching tasks, yielding direct and context specific evidence of their teaching performance. Learner teachers spend considerable amounts of time in classrooms teaching their own groups of students. It can be expected that they do so without assessors or mentors being present in the classroom, except for some formal visits.

This situation opens up many possibilities for the use of video as a means to collect evidence for developing proficiency. For the purpose of learning video registrations of teaching enable learner teachers to share their experiences with other learners teachers, mentors and assessors, which can aid the mutual process of reflection on teaching actions. For assessment purposes video can be used as a direct source of evidence of teaching performance as is takes place in the here-and-now situation of the learner teacher.

However, the use of video for assessments comes with certain challenges. Authentic and context specific evidence can at the same time be difficult to untangle for the assessor. For the student, the delivery of acceptable, perceptible, assessable evidence can be frustrated by conceptual, practical and ethical problems. A first measure to improve the assessability would be to rely not just on one filmed occasion of teaching but on registrations of multiple occasions. This is the case in what can be referred to as video portfolios (Frederiksen, Sipusic, Sherin, & Wolfe, 1998; Bakker, 2008).
A video portfolio is a structured and documented collection of video-evidence which shows the teachers' performance (Bakker, 2008). The collection

1). is aimed at illustrating, discussing, reflecting and judging the quality of a well demarcated domain of teacher competence;
2). pertains to a well defined set of critical situations, which elicit the relevant actions;
3). provides directly perceptible and essential video evidence in terms of teacher activities and related student activities, which are visible, audible, and interpretable;
4). contains necessary context information.

If video portfolios are to be composed by individual learner teachers during practical parts of their training (internship, induction period) certain measures will have to be taken to meet requirements regarding assessment and learning functions. In this paper we concentrate on the assessment requirements.

The two central questions are:

- Which design characteristics can be discerned for video portfolio that contribute to the valid assessment of teachers?
- To which extent do practical applications of video portfolio in the context of pre-service and in-service teaching meet these design characteristics.

In this paper a synopsis of four case studies is reported. Three studies were carried out in the context of in-service teachers and one within the context of pre-service teachers in the Netherlands. The purpose of the studies was to discern, develop and evaluate specifications for video portfolio assessment. For the elaboration of specifications we drew insights from four related fields of study to be described below: views on teaching and teacher education, validity issues around performance assessments, the use of video for teacher learning and performance assessment, and principles of design based research.

2. Theoretical background

2.1 Changing views on teaching and teacher education

Teacher preparation programs have been undergoing a major change towards learner-centered, context-specific and authentic learning during the last decades. The major driving force behind this is to bridge the perceived gap between what is learnt in teacher preparation programs and what is asked and needed at work (Veenman, 1994). During their preparation period in institutes for teacher training learner teachers carry out meaningful and realistic teaching tasks. Learner teachers who participate within professional development schools (PDS) even experience the full complexity of teaching within a complex and authentic environment (Darling-Hammond, 1989; Levine, 1988).

This shift in emphasis within teacher preparation programs can be attributed to a change in the view on what constitutes competent teaching. Without pretending to be complete different elements of teacher competence have been emphasised throughout the history of evaluating teachers: (a) personality traits which help to make a successful teacher (Getzels and Jackson, 1963; Creemers (1991); (b) knowledge elements involving subject matter content, ways teachers think within a discipline (Bruner, 1963; Tom and Valli, 1990); (c) forms of teacher behaviour which contribute to learning performance (Brophy and Good, 1986; Simon and Boyer, 1974); (d) teachers' cognitions and decision-making processes (Kagan, 1990; Verloop, 1988); (e) teachers' practical knowledge which they apply to specific situations in which they find themselves (their class, their subject domain) and the way they form theories about these situations (Beijaard and Verloop, 1996). f) the functions of teaching for students' learning activities and learning results, including cognitive, affective, motivational, metacognitive and developmental factors of learning (Shuell, 1993, Vermunt & Verloop, 1999; Vermunt & Verschaffel, 2000).

In contemporary teacher education, the combination of all these elements above within specific contexts is stressed. As Darling-Hammond and Snyder (2000) put it: "All teaching and all learning is shaped by the contexts in which they occur. These contexts are defined by the nature of the subject matter, the goals of instruction, the individual proclivities and understandings of learners and teachers, and the settings within which teaching and learning take place. Such variables as school organization, resources, materials, amount of time and how it is structured for learning, the duration and nature of relationships among students and teachers, community norms and values influence the processes and outcomes of teaching decisions" (p. 524).

For purposes of teacher assessment Roelofs and Sanders (2007) have recently developed an eclectic model of teacher performance which captures the elements described above, taking care of the context-specific nature of teaching. The starting point in this model, represented in Figure 1, is that teacher competence is reflected in the consequences of teachers' actions, the most important being students' learning activities. Other examples of consequences are: a (smooth or disruptive) classroom climate, a feeling of well-being among students, good relationships with parents and colleagues. Starting from the consequences, the remaining elements of the model can be mapped backwardly. First, the component 'actions' refers to professional activities, e.g., delivering instruction, providing feedback to students, and creating a cooperative classroom atmosphere. Second, any teacher activity takes place within a specific context in which a teacher has to make many decisions, on a long term basis (planning ahead) or immediately within a classroom situation (cf. Doyle, 1983). For instance, teachers will have to plan their instruction and adapt it depending on differing circumstances (e.g., different student learning styles, different organizational conditions). Third, when making decisions and performing activities, teachers will have to draw from a professional knowledge base and from some personal characteristics.

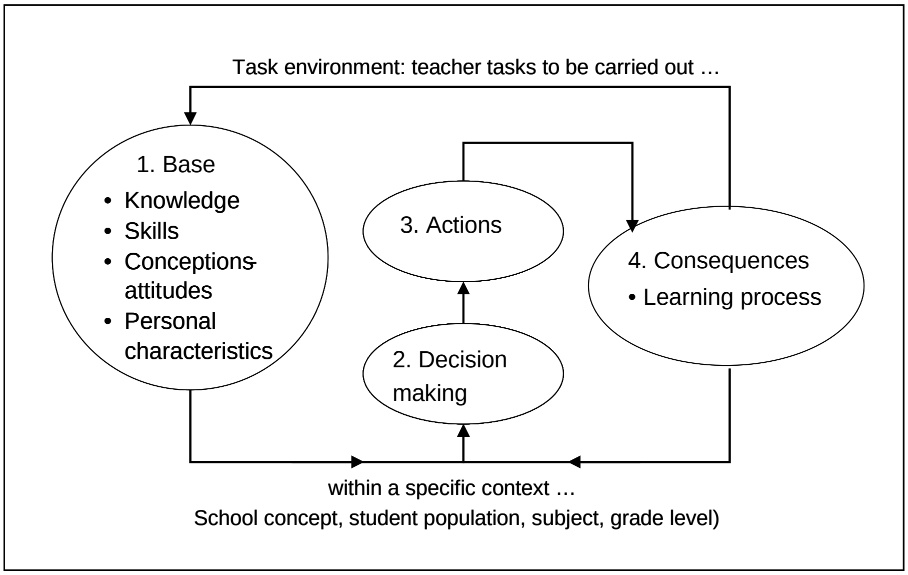

Task environment: teacher tasks to be carried out …
1. Base
• Knowledge
• Skills
• Conceptions-attitudes
• Personal characteristics
3. Actions
4. Consequences
• Learning process
2. Decision making
within a specific context …
School concept, student population, subject, grade level)

Figure 1: Process oriented model of teacher performance (Roelofs & Sanders, 2007)

A first design implication to be derived for video portfolio it that it should give an account of the full teacher performance including decisions, actions, consequences for the students. This will mean that all sources of evidence should support these aspects of teaching performance.

2.2 Performance assessments and their validity

The change in view on teaching and teacher education has resulted in a change in the view on teacher assessment. There is a shift visible from objective testing to performance assessment. It is realized that assessment practices can have a profound influence on how students study and on how their educators teach them (Crooks, 1988; Dochy, Moerkerke, & Martens, 1996; Fredricksen & Collins, 1989; Madaus, 1988). To the extent that the assessment will direct the attention of educators and student teachers to the kinds of performances included in the assessment, the assessment should involve the kinds of performances that are of most interest. Alderson and Wall (1993) and Prodromou (1995) have described this as the "backwash effect": what is assessed strongly influences what is learned. If assessment only measures factual knowledge, then learners will concentrate primarily on learning facts.

A growing recognition of the limitations in objective tests and concern about the impact of assessment on learners have generated renewed enthusiasm for performance assessments (Fredricksen, 1984; Fredricksen & Collins, 1989; Vu & Barrows, 1994; Wiggins, 1989, 1993).

The direct assessment of teacher competence seems to require that complex performances be evaluated. If we wish to evaluate student teachers abilities to design and deliver math instruction, the most direct approach to assessment would be to have them design and deliver instruction one or more math lessons. Typically, in a performance assessment, the teacher will be asked to perform, produce or create something over a sufficient duration of time to permit evaluation of either the process or the product of performance, or both (Messick, 1994).

Typically performance assessments:

- aim to be realistic representations of the actual tasks/activities we want to assess competence in;
- aim to provide meaningful and relevant tasks/activities that are worth doing;
- allow examinees substantial freedom in the interpreting of, responding to, and even designing or selecting of tasks/activities;
- take a considerable amount of time;
- require expert judgment in scoring (Gipps, 1994; Haertel, 1991; Moss, 1994; Swanson, Norman, & Linn, 1995; Wiggins, 1993)

Performance assessment however also have drawbacks. Certain issues that were largely solved by objective testing re-emerge as serious problems. The appearance of fidelity or authenticity does not necessarily imply that a proposed interpretation is valid (Messick, 1994). The scoring of complex responses requires human scorers, and therefore the consistency of scoring is a potential concern, albeit one that seems manageable (Mehrens, 1992). The problem of task specificity (low correlations between scores on different tasks), which limits the extent to which results from small samples of tasks can be generalized to broader domains, is a serious concern. The difficulty in equating different forms of performance assessments may also reduce the utility of this kind of assessment in high-stakes applications (Mehrens, 1992).

Following the line of thought an next design implication for video portfolios is that the problem of generalizability should be taken care of. In other words, if the domain of competence is clear (e.g. instruction), it should also be clear in which task context settings (e.g. small-large groups, subject areas) the learner teacher should collect his video evidence.

2.3 The use of video for teacher learning and assessment

The use of video registrations of teacher task performance holds a promise for combining the requirements of educativeness and validity in performance assessments. Earlier research revealed that student teachers can and will learn from making and discussing digital video recordings of their teaching behavior. In his overview of possible learning outcomes of teachers' interactions with video Sherin (2007) mentions four areas: 1) doing, i.e. improvement of teaching skills, 2) engaging, i.e. enhancing teachers' motivation to learn more about a topic, 3) seeing, i.e. paying attention to particular aspects of instruction and to recognize certain kinds of events as meaningful; and 4) saying, i.e. helping teachers to acquire new facts and explanations and to increase their ability to recall and converse.

Although similar methodological challenges may be expected for video registration as compared to live classroom observations in terms of the sampling design, and issues related to the evaluation system used (cf. Evertson & Green, 1986; Stodolsky, 1990) the former offers advantages above the latter (Jaeger, 1993). The most important advantage is the possibility of multi rater assessment. Different assessors can base their judgments on the same evidence, without the necessity of being present in the class. Secondly, assessors can review episodes as many times as desired before deciding to give a judgment. Thirdly, sources of rater inconsistencies can be studied in detail when reviewing episodes. Fourthly, if available, modern audio-visual technology enables high-fidelity registration of conversations, interactions between teachers, students and learning materials. In live observations, once a critical interaction is missed the evidence is lost.

As the assessor is not necessarily present during the lessons, he/she will have to rely on the eyes of the camera operator. This may cause several disadvantages of using video. Firstly, it may be difficult to have glimpses, and short overall views, normally made by live observers, without the use of multiple cameras and smart editing techniques. A second disadvantage is the dependability on the focus of the camera operator. Does he pay attention to the most important scenes in the pedagogical situation? This problem might become bigger in complex interactions compared with one-to-one tutoring. Van Es and Sherin's findings on the use of video for teacher learning suggest that it takes a learning process before student teachers are able to select critical events from their lessons as evidence for their teaching competence (Van Es & Sherin, 2006). The question is, who will operate the camera and what do we require from him/her, when the learner teacher is busy teaching?

Design implications to be derived from this field of study relate to the problem of perceptibility. First, the camera operator should get close to the actual teaching and learning process. Second, he should catch the most critical scenes of the teaching learning process.

2.4 Design based research

Since 2003 Cito has – in cooperation with educational consultants and two teacher colleges - been carrying out design studies to uncover characteristics of video portfolios that contribute both to valid interpretations and to the criteria of educativeness.

Using the method of design based research the overall goal is to bring real problems in educational practice closer to a solution. Design-based research is not so much an approach as it is a series of approaches with the intent to produce new theories, artifacts, and practices that account for and potentially impact learning and teaching in a natural setting (Barab & Squire, 2004). Design-based researchers intend to find theoretically underpinned and empirically tested design principles and methods (Reeves, 2006; Van den Akker, 1999). As is the case in design studies, we started with an analytical phase in which the knowledge and ideas about VP assessment were explored and combined. During the next phase these design characteristics were implemented within proofs of concept of video portfolio.

In the remainder of the paper the further development and implementation of design criteria for video portfolios is described in three design-based in-service instructional contexts. In the fourth study the actual practice of video portfolio use in pre-service context is evaluated against the design characteristics developed in the preceding studies.

3. Method

3.1 Research contexts and subjects

In three consecutive studies initial design characteristics of video portfolio were implemented in proofs of concepts. All proofs of concept video portfolios were judged and evaluated by trained assessors. Results of the different studies were fed back into the blueprint for VP's under construction.

Context 1 pertained to Kindergarten teachers in-service who took part in an early child development program "Pyramid', for which they had been trained. Proof of concept VP's were developed to enable assessors to give judgments about the teachers' instructional performance and the assessability of the collected evidence. The content of the early child development program aimed at helping young students to accelerate their acquisition of concepts. This was to be accomplished by instructing and scaffolding so called distancing techniques, which enable students to form progressively abstract representations, a basis for concept development (Cocking & Renninger, 1993). All instructional activities were organized around projects, each lasting two weeks(for instance 'autumn' or 'water'). Three experienced teachers from three different schools volunteered to be subject teachers in the VP. VP's were designed showing teachers' instructional performance in whole class contexts (two teachers, age 35, 47) and within tutor group contexts (one teacher, age 54). Nine teacher trainers involved in the training for using the Pyramid program participated in an interview study.

Context 2 pertained to senior secondary vocational education and involved the coaching of small groups of students working collaboratively on complex authentic tasks, by means of which they would acquire technical competence. The teachers' task was to coach the students throughout the project period in such a way that they would reach higher levels of understanding and technical performance. Four teachers in the role of coaches participated as volunteers for the proof of concept VP's. Each of them coached approximately six groups of three to four students during each project period of four weeks. A total of 6 assessors participated in the design study (Bakker, 2008), where they were asked to judge the VP's on a elaborated criteria and to review the process of judgment using the VP evidence.

Context 3 related to the field of driver training. As part of innovated driver instructor exams, candidates will be enabled to construct own VP's during their internship period. To prepare for this new practice proof of concept VP's were developed to evaluate the feasibility of VP practice. Three driver instructors volunteered in the study, two of them beginning instructors (male, age 24 and 26, one year experience), two of them experienced (both male, age 55 and 60 respectively with 17 and 20 year experience). Three assessors worked at a national examination institute, responsible for the instructors' exams. All were experienced, aged 40, 61, 61 with 5, 14 and 14 years of experience respectively.

Context 4 related to teacher preparation at university teacher colleges. Four teacher colleges participated in the study. At all colleges learner teachers had a degree in one or two school subjects. During the teacher preparation stage they were trained to be teachers in one year. All learner teachers were enrolled in full internships at schools where they had a partial responsibility for delivery of classes at general secondary education. During this period they were also expected to collect video evidence of their teaching performance. In doing so, they made use of a web-based digital learning environment in which integrated web-tools were available for uploading, and editing video footage, and uploading supportive evidence (teacher reflections, lesson designs). Also an exchange tool was built in, which enabled teacher trainers to give specific assignments and fellow-learner teachers and mentors to give feedback on teaching performance and learner teacher reflections. The project was referred to as ‘DIVI-dossiers’ and was funded by the SURF-foundation, a collaborative organization for higher education institutions aimed at innovations in ICT.

Although the number of participating teacher colleges and learner teachers was much larger, for logistic reasons four locations were chosen which were considered ahead in terms of video use for educative and assessment purposes. 11 learner teachers (6 male, 5 female, mean age 25.4 years) were asked to participate in the study of their individual VP’s. The distribution over colleges and subjects is displayed in Table 1.

Table 1 overview of participating teachers from different teacher colleges and teaching different subjects

College	Physics	Chemistry	History	Geography	Biology	Spanish	French	Total
E	3							3
A		1		1	1			3
G			2	1				3
J						1	1	2
Total	3	1	2	2	1	1	1	11

Additionally four associated teacher mentors from the teacher colleges, who played crucial roles in the implementation of the project DIVI-dossiers participated in interviews on the use of VP’s for assessment purposes (two male, two female ages respectively 32, 48, 55, 35).

3.2 Development of a general design framework

Following the method of design based research, design characteristics for a video portfolio procedure were derived from the fields of study described above. The focus was on the development of a blue print for a procedure which yields assessable video portfolios, resulting in valid interpretations about teacher task performance. The overall framework was drawn from the literature on validity arguments.

Kane (2004) recently proposed an argument-based approach to validity by which performance assessments can be validated. In his argument-based approach, Kane contends that the validity of an assessment can be studied by evaluating the chain of inferences that takes place when interpreting its outcomes. More specifically four inferences form the heart of the validity argument: (1) evaluation of observed performance, yielding an observed score. (2) generalization of the observed score to the expected score over the assessment domain, (3) extrapolation from the assessment domain to the competence domain, 4) extrapolation from the competence domain to the practice domain. Depending on the assessment purpose (e.g. formative or summative) a specific fifth inference is applicable: decision about readiness for practice or follow-up learning activities to be undertaken.

The validity of the interpretation assigned to the results of video portfolios depends on the plausibility of the inferences leading from the observed performance to the target score.
Kane’s approach served as a general design framework for the three proof of concept video portfolio studies as far as the assessment function concerned. Insights from the fields of study described above were used to elaborate design specifications in every design stage. For the *design* of PA’s the chain of inferences starts where the interpretation of scores ends, by elaborating a domain of competence from the practice domain. In designing PA’s we need a chain of design interpretations starting from a complex practice domain and ending up in the collection of task performances (see figure 2).

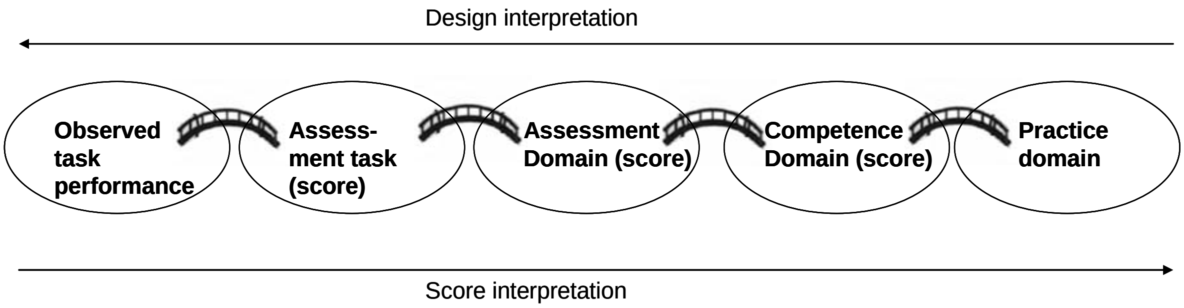


Figure 2: Design and score interpretation chains as inverse activities (adapted from Kane, Crooks & Cohen, 1999).

3.3 Initial design characteristics

Design step1: elaboration of competence domains
To warrant any extrapolation towards the practice domain of teaching it is needed as a first design step that teacher activities are elaborated in a domain of competence which play a major role role in practice. For the three studies research on the relevant domain of competence was reviewed. This pertained to the repsective domains of concept acquisition for young students, coaching in vocational education, and driver training. For elaborating the content of the competence domain a research base of basic and applied research, empirical practice analyses was used, including empirical studies of the relationship between competence in various domain aspects activities and practice outcomes. Panels of experts were consulted to check and revise the domain of competence. The general model of teaching performance served as a process model, an important vehicle to describe the processes teachers engage in when interacting with students for learning.

Context 1:
In the K-context a domain of competence was elaborated based the tenets in the early child development program ‘Piramid’ (see Van Kuyk, 2000). This program aims to stimulate the development of young students and where possible to accelerate it. A central principle is that students make a next step in their development of concept mastery when they can distance themselves from the here-and-now, a process that is also known as ‘distancing’ (Cocking & Renninger, 1993).
A framework was constructed in which teacher interventions were elaborated that are expected to contribute to concept acquisition. Four stages of acquisition were discerned and within each of them stage relevant interventions for concept acquisition were elaborated, connected with expected influence on student learning: orientation, initial acquisition, broadening understanding, and acquiring deep understanding.

Context 2:
Within the vocational education context coaching was the central domain of competence. It was elaborated as stimulating and supporting self-regulated learning (SRL; Boekaerts, 1999; Boekaerts & Simons, 1995; Bolhuis, 2000; Butler & Winne, 1995). Typical coaching interventions which have a known influence on SRL include asking questions and providing feedback (clues, hints, advice, and examples) on learning activities employed by students. The elaboration of learning activities related to cognitive, affective, meta-cognitive and collaborative learning. Cognitive learning activities refer to activities to process subject matter and that lead to learning outcomes in terms of changes in students' knowledge base and skills (cognitive learning activities). Affective learning activities pertain to coping with emotions that arise during learning and that lead to a mood that fosters or impairs the progress of the learning process (affective learning activities). meta-cognitive learning activities concern thinking activities that students use to decide on learning contents, to exert control over their processing and affective activities, and to regulate the course and outcomes of their learning (meta-cognitive learning activities). Finally, learning activities that pertain to collaboration with other students. The knowledge about coaching for self-regulated learning, encompassing the first three learning activities mentioned was based on instructional theories as elaborated by Shuell (1993), Vermunt and Verloop (1999), and Winne and Hadwin (1998). Coaching on the fourth learning activity is based on theories about collaborative learning (Johnson and Johnson, 1994; Slavin, 1990).

Context 3:
In the (practical) driver training context the one-to-one learning situation is typical, including a mix of individual instruction and coaching activities. The domain of competence was elaborated employing principles of coaching and instruction as described above but then linked to a view on the acquisition of driving proficiency. According to this view learner drivers acquire skills on three levels: 1) vehicle control, including all kinds of manoeuvres, 2) traffic participation in a) simple and b) complex situations, including all kinds of traffic tasks (e.g. merging, crossing), and strategic tasks, like preparing for driving, choosing or changing a route. Driver instructors are expected to help advance learner drivers from the level of vehicle control towards the level of strategic driving, where lower levels are automated.

The instruction domain entailed the following interventions: introduction (lesson overview, motivating, connection to prior learning, learning objectives), instruction (content, discussing task components, correcting errors), guided practice (places to practice, increasing complexity levels, intensive practice, enable errors within safety margins), feedback (rounding up the lesson, discussion of learner results and learning process).

The coaching interventions entailed: monitoring driving task performance (detecting driver errors, thinking aloud), supporting the acquisition process (scaffolding and correcting, advice for improvement, explanatory feedback, timing), adaptation to the needs of the learner (level of performance, level of independence), fostering positive learning intention (complimenting, acceptance of errors, positive expectations in learner), creating positive learning climate (making contact, level of language use).

Design step 2: elaboration of the assessment domain

The *second* design stage relates to the way the comptence domain is elaborated into an assessment domain, i.e. the universe of possible assessment tasks on which inferences about the teacher competence are based.

Based on the general process model of teaching performance the characteristics of assessment tasks in which competent action is to be shown were elaborated. The video portfolio was to show a teacher who acts in a real learning situation (classroom, group, car), using a repertoire of content specific acitivites, by which he deliberately sets out learning activities on the part of the students. Using the process model it became clear that teacher decision making would call for supportive evidence not to be covered by the video camera. Additionally, context information would be necessary to be able to fully understand the decisions and actions taken by the teacher (Table 2).

Table 2: Basic parts of video portfolio related to components of instructional performance

Parts of video portfolio	**Evidence source**	**Components of performance**		
		Decision-making behavior	**Actions**	**Consequences for students**
1. Video registration				
a Registration of video images of teacher	Collection of synchronously mixed video episodes (researcher)		X	
b Registration of video images of student (synchronous with teacher)				X
c Table of contents of filmed video episodes	Registration form of video fragments (researcher)		X	X
2. Documentation of the teaching situations				
a. Lesson artifacts and learning environment (educational tools, hand-outs.)	Concrete material and digital photos (researcher and instructor jointly)		X	X
b. Description of lesson context, plan and lesson scenario	Registration form (instructor)			
3. Teacher reflection on	Stimulated recall of	X		

interventions	interview and analysis of protocol (researcher and instructor jointly)			

Additionally, decided was on the range of tasks and task conditions, the types of evidence needed in it, thus demarcating the boundaries of the assessment domain. The task conditions included critical situations in which relevant teacher activities were expected to take place.

Within the Kindergarten context the most important conditions for gathering evidence for teacher instructional performance were the four discerned stages of concept acquisition (Cocking & Renninger, 1993) as elaborated in the competence domain, the target group (whole class all students-small groups: students at risk). Within the vocational education context the following set of conditions was specified: stage of the project (first, second, third and fourth week), group composition, a limitation to first year students (grade as fixed condition), limitation to one group of students to be coached (groups as fixed condition). Within the driver training context the conditions were: learner stage (starting, intermediate, and advanced learner); limitation to one driver training method (Driver Training Stepwise‘ (DTS), Nägele & Vissers, 2003).

Design step 3: elaboration of assessment tasks

The *third* design stage pertains to the elaboration of the actual assessment tasks and the way in which they can be carried out in classroom practice. The main design question is: what, when, where and how to acquire video evidence and supporting evidence. As this involves the actual construction of video portfolio, more attention is paid to this stage. This comprises the actual assessment task, from which the results will be judged by assessors. The design principles used enable the scoring and generalization of observations of task performance towards the assessment domain.

First, to an extent possible, within all three design study contexts the video portfolio assessment procedure was standardized. This was done by selecting a fairly representative set of tasks and conditions of observation based on the decisions regarding the assessment domain as described above. In the K-context teachers were filmed during their regular project-based lessons for young students during three planned stages of concept acquisition (elementary acquisition, broadening understanding, deepening understanding). Teachers did not get any special instructions, except for realizing their intended activity plan as much as possible. In the vocational education context four meeting were scheduled across the four weeks of a project-based learning period on which the camera team would be present to film the regular coaching meeting between the four teachers and their group of students.

In the same way the four driver instructors were asked to deliver lessons to three learner drivers enrolled in different stages of training (vehicle control stage, driving in simple traffic situations, driving in complex traffic situations). No specific instructional other than to give regular lessons were given.

Second, a basic design characteristic was that all pieces of evidence for instructional performance should relate to the *same* set of teaching situations and therefore be collected in coherence. What was to be recorded comprised the teacher's actions, his/her underlying decision-making process, the observed consequences for students and the lesson context (situation) in which the teaching actions took place. In order to construct the video portfolio, task protocols were designed for 1) collecting, registering, editing, mixing and describing video footage, 2) documenting the relevant teaching situations and 3) collecting information about the teacher’s decision making process during her interventions.

In the three design contexts the video portfolios were constructed by researchers in cooperation with specially selected teachers. Every video portfolio consisted of three parts.

The first part of the video portfolio is a video registration of lesson episodes collected using a specific recording protocol based on the boundaries set for the assessment domain in terms of conditions of observation, critical situations for specific instructional activities procedures. In all contexts at least two DV cameras were used. One fixed DV camera was directed at the teacher and one hand-held camera was directed at the students to catch their learning activities.

Camera operators were informed about the instructional plans of the teachers by the researchers. In addition they were instructed to stay close to the scenes: to walk to the students and take detailed shots of their interaction with learning materials and with peers. All teacher activities were to be followed closely, including writing on the blackboard (in whole group settings), pointing at pictures. In the driver training context a choice was made to attach a webcam on the front window to capture the traffic situation in front of driver and instructor. The other webcam was directed at the learner driver. Both webcams were connected to a tablet pc, with specialized software which resulted in automatic synchronization and compressed video files. Unlike the procedure used in the Kindergarten and the vocational context the instructor was audible but not visible. In all three contexts extra microphones were placed to capture all student responses.

According to the design principles of the assessment domain the basic assessment task would consist of delivering documented video episodes of meaningful instruction/coaching situations. An episode is a period of time within which meaningful and discernable instructional activities were carried out: previewing the day, explaining the meaning of an activity, discussing. To end up with these basic assessment tasks three measures were taken. First after recording, the lesson or meeting was divided into meaningful episodes according to the descriptions above. Second, the content of these episodes in terms of teacher and students activities was described. Next, filmed episodes were selected that would provide relevant evidence of instructional or coaching competence. These pertained to the activities that were considered critical in the assessment domain (e.g. opening the lesson, preview, start of discussion of students' work). Third, on average six episodes out of each lesson were selected on two criteria: perceptibility (visibility and audibility) and relevance for judgment on instructional or coaching performance. Finally, the teacher oriented (or road-oriented) footage and the student oriented footage were synchronized and mixed, using professional software (AVID Pro; Adobe Premiere). In the driver training context this synchronizing occurred automatically.

Except for the driver training context the view on the teacher was taken as a starting point. During moments when students answered questions or carried out tasks, the view would change to the students, to ensure that their reactions to teacher actions would be visible. When student responses changed quickly from one student to another a group picture was used.

A second part of the portfolios consisted of a documentation of the teaching situations. These entailed pictures of lesson setups (e.g. arrangements of furniture, the learning environment) and (play-) learning materials which were used during the videotaped lessons. Information was added about the current instructional unit (project, stage of learning), its contents, the planned learning activities, what went before, what would follow. Reference was made to specific passages in teacher manuals or the lesson plan (in case of driver training).
A third part of the VP, which was not added in the driver training VP, is a report of underlying decision-making behavior, obtained by means of a 'stimulated recall' interview on the basis of a one track video registration, shortly after the registration. In these interviews, also recorded on video, a situation specific reflection is asked of teachers. To prepare for this interview, teachers received a videotape with a registration of all 6 selected lesson episodes. The tape was accompanied by a preparation form for the interview. For each lesson episode, questions were asked about considerations for the selected interventions. Teachers indicate why they had undertaken certain interventions, which effects they themselves observed, and how they would justify their actions based on pedagogical-didactical principles. Some recurrent reflection questions in the interview were: what did you do in this episode? what were the consequences of that for students? To what extent did you adjust your approach to that? What were your motives/considerations to act like that? What did you think while you acted like that? Which characteristics of the situation were important (like class, student characteristics, accidental situation). In this manner an image is reconstructed of the teachers' considerations prior to and during the filmed episodes.
The driver instructors had not been interviewed, because this did not fit into the existing practice of instructor examination and had met resistance from the instructors' branche of trade.

3.4 Data collection

Context 1
In the K-context an interview study was conducted. Two main topics were addressed in the interview. First, respondents were asked to value the quality of the instruction provided by the teacher of whom the video portfolios hade been produced. Second, they were asked to rate the quality of the evidence collected and presented in the video portfolio in terms of perceptibility, representativeness of content

and processes. Finally, they were asked to indicate whether information was lacking that was needed to rate the quality of instruction.

Context 2
In the vocational education context six trained assessors scored the four video portfolios constructed for this purpose. Each assessor scored three of the four video portfolios, because scoring of all the portfolios would have taken too much time. The assessors installed the video-data base application (MILE) containing the video portfolios on their own computers, and scored the video portfolios independently and at their own pace. After scoring the video portfolios individually, the assessors met in pairs to discuss the scored portfolios.
To obtain more detailed information about factors that stimulated or hindered the assessors in making valid interpretations and judgments, they were interviewed. After scoring the three video portfolios, all assessors participated in a semi-structured interview about their experiences with the composition of the video portfolios, and with the interpretation and judgment of the individual video episodes and overall video portfolios.

Context 3
In the driver training context an examiner was present during each of the three lessons that would also be part of the video portfolio. During and immediately after the live visit he scored the lessons on all performance criteria form. At the same time a researcher was present with a laptop which automatically registered the lesson as described above for the purpose on forming a VP. The complete VP consisting of 3 lessons was scored by all assessors, using for each instructor a video CD with the three registered lessons, supported by lesson context data. Also in this context the assessors were interviewed. The topics of the interview pertained to the perceptibility of the video evidence: the extent to which learner drivers' actions and the interaction between instructor and learner driver were perceptible and how this affected their scoring behavior. Also the practical feasibility of VP assessment was evaluated in terms of time needed to arrive at an overall judgment compared with live observations.

Context 4
Because in the teacher education context learner teachers themselves constructed the VP's the produced VP's were analyzed by the researchers scoring them on the presence of the design characteristics which were used in the first three proof-of-concept studies. The scoring took place by using a system with 15 entries, of which nine are predominantly descriptive in nature and 6 evaluative in nature: 1) From which perspective is the lesson recorded (students, group or learner teacher as main picture?). 2) Which camera position was used and was there a camera operator? 3) In which organizational setting did the learner teacher and the students work? 4) Is the teacher visible and audible (when headphones are used for viewing the VP)? 5) Are students and their interactions with teacher, peers and learning materials visible and audible? Are the media used by the (learner) teacher visible: 6) Blackboard and 7) other learning media (working books, objects)? 7) Was a lesson plan available in the VP? 8) What was the theme of the lesson? 9) What subject was it about? 11) Is a description of the instructional context available?
12) Are (references to) learning materials available? 13 Is information available about teacher decision making?To which extent is the situation relevant for the competence domain to be scored in terms of 14) the processes covered and the 15) the critical situations to which is generalized?

To check the results of the coding the four mentors of the learner teacher were interviewed.

4. Results

Context 1: Instruction in Kindergarten
The construction process resulted in one DVD per teacher containing selected episodes, descriptions of the lesson activities, underlying information about the decisions process as was planned (see table 2). Additional paper booklets with all supportive information was made available as well. Teacher consultants could view the three DVD's by clicking hyperlinks for different sources of evidence or by reading the booklets. Additionally, they were provided with an assessment framework to give an

overall rating of the quality of the instruction. The major part of the study was the evaluation of the quality of the evidence encountered in the VP.

In Table 3 descriptive and evaluative characteristics of the VP’s are displayed. In the interviews teacher consultants rated the perceptibility of the video episodes positively. They could clearly see, hear and understand what was going on during the instruction to small groups of young students (aged 4 up to 6) and to al lesser extent to what was going on in whole group contexts.

They mentioned specific measures which improved the perceptibility of what students said in response to teachers, among which zooming in into individual students and using a picture of the students working as the main picture and incidentally switch to the teacher instead of vice versa.

According to the consultants some situations were better perceivable than others. First these involved situations with lively interaction between teacher and student and between students. Second, it pertained to situations in which teacher and students use concrete learning materials, e.g. the teacher asks questions while pointing to spots on a picture; having students point at a picture (see figure 2); letting students feel objects and have them name them; sorting garments by the season in which they are worn.

As far as the representation of the whole process of teacher performance (decisions, action, students’ consequential activities) concerned they had some doubts. More specifically they doubted whether they fully understood the actual consequences in terms of learning activities undertaken by the students. In case of whole class situations observations about the learning process of individual students were hard to make. This would also pose difficulties in rating the quality of the teachers’ instruction. Statements about the teachers’ contribution to students acquisition of concepts were mostly based on what was observed on the part of the teacher

’What she doesn’t show is that she, when focusing on children’s vocabulary, should repeat the words at least a couple of times and in different ways.

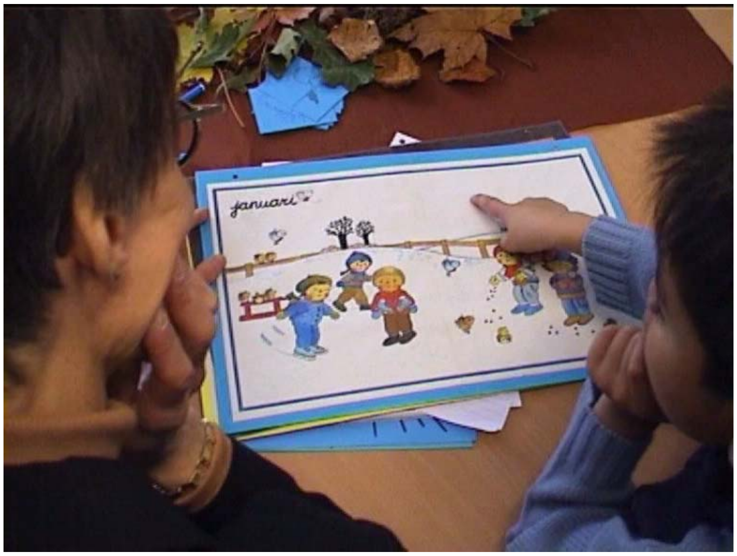


Figure 2: Zooming in to improve the perceptibility of interaction with learning materials

Context 2: Coaching in vocational education
The professionally constructed VP’s in the vocational education design study appeared in the form of a documented video-database. It consisted of video episodes, their content descriptions, evidence regarding the coaching context (school, groups, student characteristics, learning material), teachers’ decision making process, and students perceptions of teacher coaching behavior (see figure 3).

The six assessors involved evaluated the evidence quality of VP in the context of the assessment of coaching. Detailed results of this study are described in Bakker (2008) and in Bakker et al (2009). For the purposes of this paper we only describe direct evaluations regarding the quality of the evidence within the VP. The results are summarized in table 4.

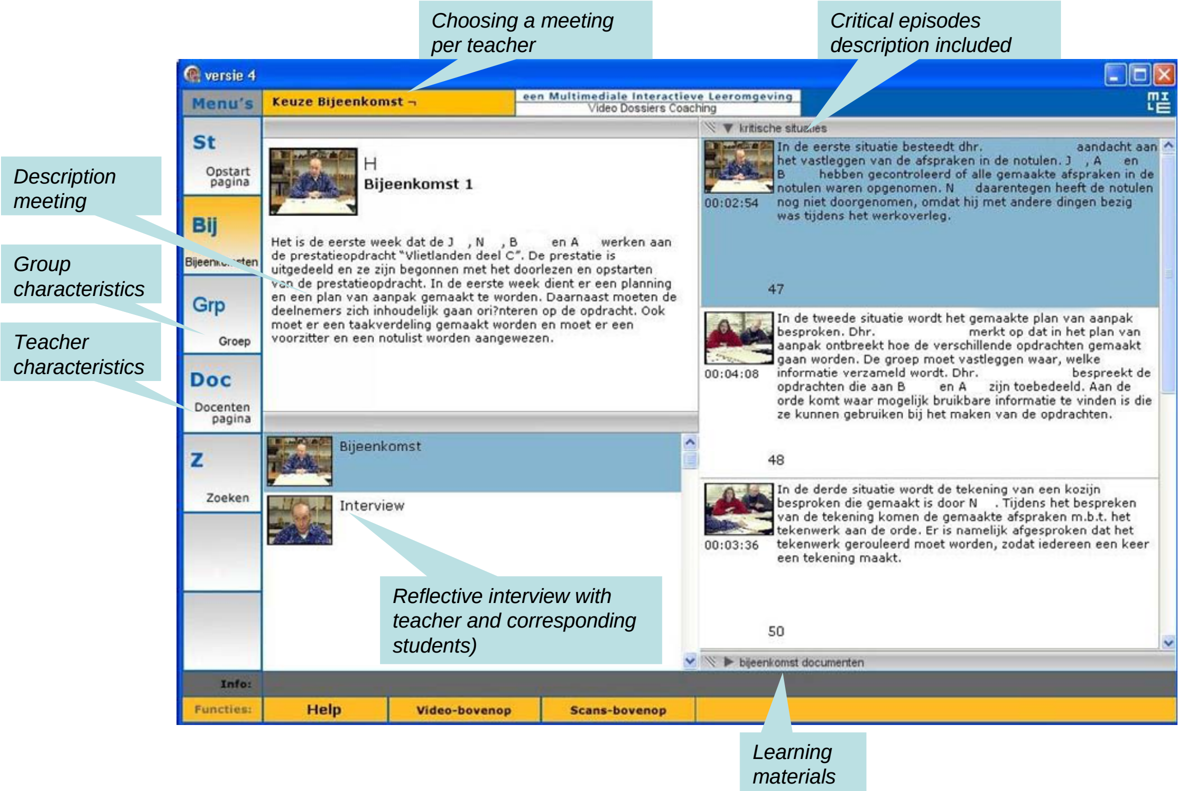
As the VP had been constructed in cooperation with a professional production company, special attention had been given to the perceptibility of teacher and students behavior as shown on the video footage. Three separate microphones had been used to catch all necessary sounds. Three synchronously mixed video tracks were available to cover any changes in interaction between students and teacher. These technical arrangements in a special video room in which only one group of students was present with its coach, resulted in highly visible and audible coaching scenes. The assessors confirmed this perceptibility.
The editing of the video episodes, based on selections made by the researcher, resulted in an even better perceptibility: close-ups from technical drawings, non-verbal cues in the faces of students and teachers, which was again confirmed by the assessors.

Nevertheless, during the interviews it appeared that although perceptibility was at a high level, some of the assessors experienced problems in understanding the use of extensive technical terminology. To be able to give a judgment on the quality of coaching for subject related cognitive learning activities it was necessary to understand details of the conversations about some technological design. Problem raising examples were conversations about ‘clamping pins’ and ‘spring bolts’.

With regard to the process representation of the VP – are all aspects of professional performance according to our proposed model - the assessors posited that all sources of information added to the judgment of coaching quality. For the most part assessors used the video episodes, and their content summaries, the interviews with teachers and students’ and the students’ background information in scoring and judging the video portfolios.
The interview with the teacher was used to find out what his aims and intentions were for each critical situation. The content summaries aided to focus attention to the relevant aspects of the coaching scenes: the aspects of performance and perceivable consequences for the students. Interviews with students gave some information about the possible consequences for the students. A shortcoming in the eyes of the assessors was that not all students involved in the coaching sessions had been interviewed, so they could not determine whether the coaching had been effective or ineffective for all students. Additionally, assessors suspected social desirable answers form the students regarding their coaches.

As regards the representativeness of the evidence collection the assessors indicated that, after having viewed (all) six video episodes from one teacher, they had developed a clear view on the teachers’ overall coaching competence. Assessors found it hard to evaluate teachers’ contributions to student learning on the basis of single video episodes.
Although most of the episodes lasted between five and ten minutes, in some cases video episodes lasted a long as 14 minutes, because a specific learning need had to be addressed by the coach. In those cases assessors had problems to keep their attention on the coaching episode. Assessors indicated that video episodes of five to ten minutes provided enough information on teachers’ performance in that situation.

Finally, the assessability of individual video episodes varied across different characteristics:
Easier to judge were episodes, in which the teacher’s behavior corresponded to his intentions as explained in the reflective interview. Another factor was the overtness or clarity of the learning need of the student: was there a problem in self-regulation, did he/she have some misconception, did he have problems in motivating himself to act? In case of multidimensional problems on the part of the students it was more complicated to assess the teacher’s behavior as a coach.



Choosing a meeting per teacher

Critical episodes description included

Description meeting

Group characteristics

Teacher characteristics

Reflective interview with teacher and corresponding students)

Learning materials

H
Bijeenkomst 1

Het is de eerste week dat de J , N , B en A werken aan de prestatieopdracht "Vlietlanden deel C". De prestatie is uitgedeeld en ze zijn begonnen met het doorlezen en opstarten van de prestatieopdracht. In de eerste week dient er een planning en een plan van aanpak gemaakt te worden. Daarnaast moeten de deelnemers zich inhoudelijk gaan ori?nteren op de opdracht. Ook moet er een taakverdeling gemaakt worden en moet er een voorzitter en een notulist worden aangewezen.

00:02:54
In de eerste situatie besteedt dhr. aandacht aan het vastleggen van de afspraken in de notulen. J , A en B hebben gecontroleerd of alle gemaakte afspraken in de notulen waren opgenomen. N daarentegen heeft de notulen nog niet doorgenomen, omdat hij met andere dingen bezig was tijdens het werkoverleg.
47

00:04:08
In de tweede situatie wordt het gemaakte plan van aanpak besproken. Dhr. merkt op dat in het plan van aanpak ontbreekt hoe de verschillende opdrachten gemaakt gaan worden. De groep moet vastleggen waar, welke informatie verzameld wordt. Dhr. bespreekt de opdrachten die aan B en A zijn toebedeeld. Aan de orde komt waar mogelijk bruikbare informatie te vinden is die ze kunnen gebruiken bij het maken van de opdrachten.
48

00:03:36
In de derde situatie wordt de tekening van een kozijn besproken die gemaakt is door N . Tijdens het bespreken van de tekening komen de gemaakte afspraken m.b.t. het tekenwerk aan de orde. Er is namelijk afgesproken dat het tekenwerk gerouleerd moet worden, zodat iedereen een keer een tekening maakt.
50

Figure 3: a screen from the video database application

Context 3: Driver training

The third design context of VP, driver training, differed in some important respects from the first two contexts. There were no camera operators, but two automatically operated webcams attached to a tablet pc. Second, the driver instructors had not been interviewed about their decision making process, due to appointments already made. This would automatically have consequences for the evaluation of the degree in which performance processes mirror the intended processes as elaborated in the competence domain.

Although the study addressed a complete evaluation of the experiments of driver instructor assessment (for details see Roelofs, Vissers & Harms, 2007), in this paper we concentrate on the evaluation of the evidence.

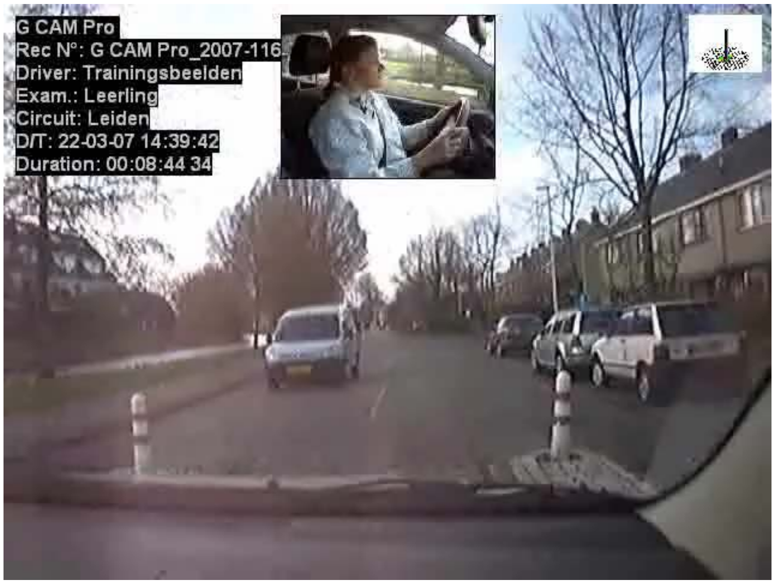


Figure 4: Double web-cam registration of driver behavior

When judging the quality of coaching by means of the video evidence the three examiners considered the interactions between the instructor and the learner driver to be well perceivable. In most cases this involved action-reaction images: the learner driver performed a certain driving task (e.g. merging, parking) and the instructor commented upon it or provided a scaffold just before the maneuver was going to be carried out.

More specifically, examiners considered the following types of situations as well perceptible: 1) verbal interactions between learner driver and instructor (question and answer; learner action and instructor reaction); 2) Traffic situations in which multiple participants are involved, multiple encounters. 3) The performance of a traffic task, starting from static situations (crossing a junction once the traffic lights switch to green). Oppositely, difficult to perceive were situations in which: 1) the view from the side mirrors was needed to judge adequacy of the instructors' response (specially in case of merging or taking right turns); 2) the physical sensation caused by speed choice and vehicle control (accelerate, braking, jerky movements) is missing. Therefore coaching aimed at improving vehicle control was considered difficult to judge. In this respect the examiners preferred live assessment above video portfolio assessment. Interestingly, the examiners were not hindered by the absence of the view on the instructors' face in assessing the instructors. The verbal interaction was considered sufficient.

The extent to which the evidence gave a full representation of the intended performance process gave a different picture compared to the previous two design contexts. The same holds for the representativeness of the chosen set of instructional episodes.

The driving instructors did not appear to carry out specific lesson plans like the K-teachers and V-coaches did. Learning objectives emerged during the driving lessons. The route taken at the particular day largely determined the lesson content, despite the stepwise learning program they were supposed to deliver.

The apparent lack of task directions appeared to complicate the interpretation of the instructors' instruction performance as the examiners noted later during the interviews. The problem was less severe regarding the interpretation of the coaching performance.

The set of instruction episodes chosen by the researcher to be judged appeared to be confusing for the examiners. The mismatch between what was expected and what was realized during driving instruction posed problems for the examiners. As a result, the examiners decided to view the complete recorded driving lessons (12 in total). They considered the chosen instruction episodes not to be representative of the complete lesson. The lack of standardization in what was to be taught during the driving lesson probably caused the actual found variations in judgments across examiners (Roelofs et al., 2007). This problem held equally for live in-car judgments. Examiners felt disoriented during their lesson visits and during their viewing of video portfolios.

As noted the problem was less urgent in evaluating the coaching performance. Examiners even considered short episodes (e.g. in which certain manoeuvres were commented upon) well assessable.

According to the examiners the lack of any information about the instructors' intentions and decisions during the lessons will have played a confusing role/

Context 4: General junior secondary education
Context four did not relate to proof-of-concept video portfolios. Instead, the VP's were produced by the learner teachers themselves. The way the assessment tasks were elaborated was one of the topics of the interview study.

Table 6 shows the elaboration of assessment tasks for students as a basis for their video portfolios.

In three out of four teacher training locations (A, E, J) the assignments for the learner teachers were rather detailed. Learner teachers are expected to show their competence in elaborated domains of competence (e.g. interpersonal of subject pedagogical competence). At teacher training location J the assignment almost has the degree of specifity as used in the proof of concept VP's in-service teacher contexts: an assignment to demonstrate competence on a specific subject pedagogical domain in a specific lesson about a specific topic for which the learner ha already prepared a lesson design. Students are asked to use a maximum of 4 video clips for one lesson. In location E Learner teachers were involved in a very specific aspect of the domain of competence 'guiding learning processes': the detection of misconceptions in carrying out a learning experiment for physics.

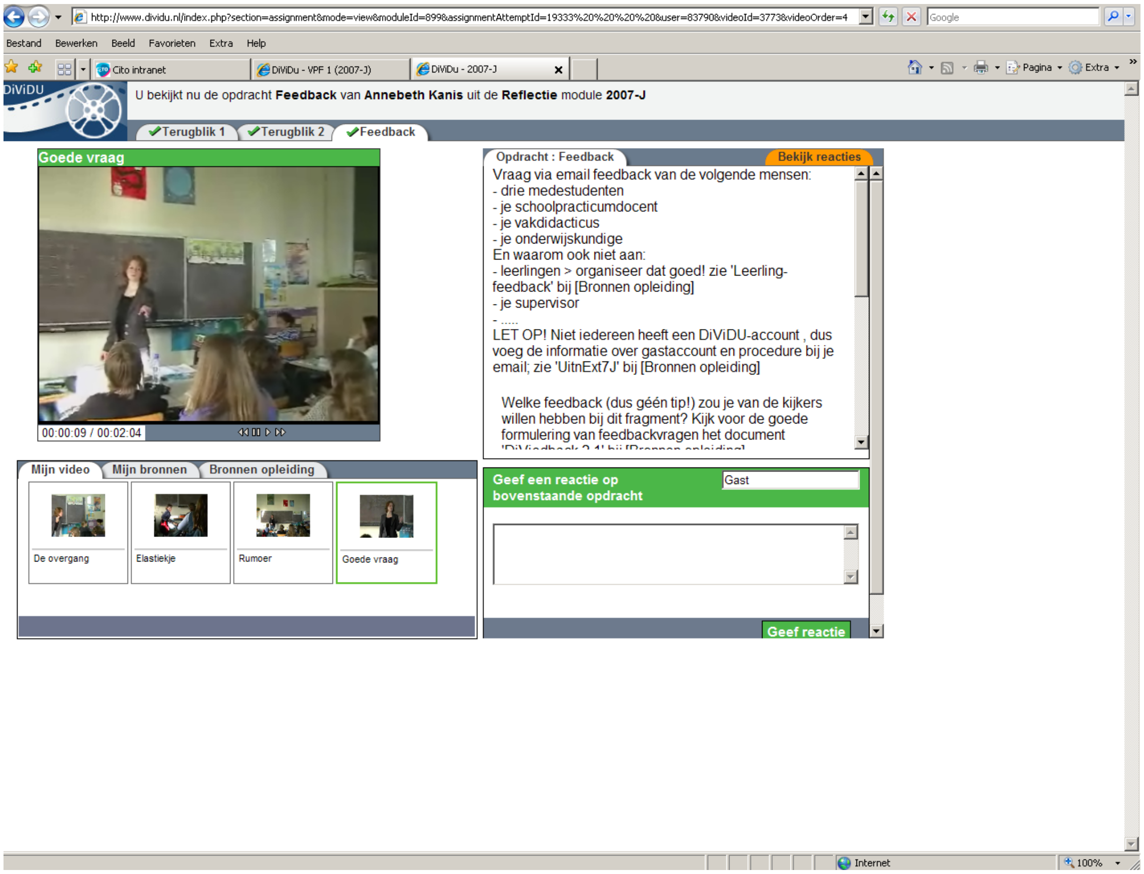
In general, the learner teachers involved did not receive detailed instructions for collecting perceptible video footage. An exception to this was location E. LT's from this location received instructions which made it necessary to film overt learning behaviors and use very concrete material. Literally: "devise an experiment in such a way that it enables students to move and grab objects to make their conceptions clear. The students are taken out of their classrooms. Choose intermediate level students or a little bit weaker. - Make boys and girls only dyds of equal level, to ensure equal participation in tasks."

At location A the assignment was less specific compared with the assignments on the other locations.

Table 7 shows the characteristics and observations regarding the video portfolios at the four locations involved. The following observations could be made:

- Video portfolios often lacked visual introductions for the viewer: where are we in the classroom? At which point of the learning process do we step in? Perceptible video evidence used the kinds of introductions.
- The perceptibility of video footage varies widely.
- Teachers are often chosen as the primary source. Therefore their actions are often perceptible. Students' activities however are less perceptible.
- In case of subject pedagogical video portfolios the evidence of helping students advance through the learning material was hardly visible. let alone interpretable. The interaction with the learning material was not filmed in detail.
- Pictures which were more perceptible pertained to rather tangible activities like: small groups of cooperating students, detailed pictures of concrete objects handled by students (ingredients for a meal; a technical drawing, a ball to be thrown).
- Hindering elements for perceptibility were: poor readability of writing on white-board or black-board, audibility of students' responses during whole class situations.

[to be completed]



U bekijkt nu de opdracht **Feedback** van **Annebeth Kanis** uit de **Reflectie** module **2007-J**

Vraag via email feedback van de volgende mensen:
- drie medestudenten
- je schoolpracticumdocent
- je vakdidacticus
- je onderwijskundige

En waarom ook niet aan:
- leerlingen > organiseer dat goed! zie 'Leerling-feedback' bij [Bronnen opleiding]
- je supervisor
- .....

LET OP! Niet iedereen heeft een DiViDU-account , dus voeg de informatie over gastaccount en procedure bij je email; zie 'UitnExt7J' bij [Bronnen opleiding]

Welke feedback (dus géén tip!) zou je van de kijkers willen hebben bij dit fragment? Kijk voor de goede formulering van feedbackvragen het document 'DiViedback 2.1' bij [Bronnen opleiding]

Figure 5: Web-environment in context 2 to upload components of video portfolios

5. Discussion

In this paper we intended to develop design characteristics for video portfolios, for assessments purposes in the context of teacher training. As an overall framework we used the validity argument approach, which was reversed in order to display the design argument for valid performance assessments. The content of the specific design steps was derived from relevant field of study, including views on teaching and teacher education, the use of video, and domain specific understanding (driver training, early child development, coaching in vocational education).

In three design contexts initial design characteristics were developed and implemented in proof of concept video portfolios. The resulting VP's were evaluated in terms of the perceptibility and relevance of their evidence.

[to be completed]

Practical implications for design of video portfolios

A practical aim of the studies was to arrive at a set of design characteristics which would help teacher training institutes to use VP as a method of performance assessment that yields both valid results and contributes to learning.

In Table 8 the first final draft design characteristics are summarized, which are expected to contribute to the aims mentioned. Most of them were result of the design studies, but some of them have emerged during the study in the teacher training context. The latter are referred to as 'new'.

Table 8: Final design characteristics of video portfolios

	Valid interpretations and decisions				**Educativeness**
	Scoring inference (observed score)	Generalization inference (Assessment domain score)	Extrapolation inference 1 (Competence domain score)	Extrapolation inference 2 (Target domain score)	
Assessment tasks					
Specific instructions for learner teacher (new)					
Some degree of standardization	X				
Relatively short	X	X			
Representative tasks		X			
Representative set of conditions of observation		X			
Challenging and complex				X	X
Performance criteria					
Based on process model of teaching and learning			X		X
Subject specific student learning		X		X	
Stage specific performance standards					X
Registration of video evidence and context information					
Teacher activities and student activities, interactions and media perceptible	X		X		X
Relevance for teaching process	X		X		X
Clarity of teaching context	X				
Connectedness of supporting evidence sources	X		X		X
Students´ own initiative					X

[to be completed]

References

Alderson, J.C., & Wall, D. (1993). Does washback exist? *Applied Linguistics, 14,* 115-129.

Bakker, 2008. *Design and evaluation of video portfolios. Reliability, generalizability, and validity of an authentic performance assessment for teachers*. Dissertation. Leiden: Leiden University Graduate School of Teaching

Barab, S. A. & Squire, K. D. (2004). Design-Based Research: Putting Our Stake in the Ground. *Journal of the Learning Sciences*, *13*(1), 1-14.

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

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.

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

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

Brophy, J., & Good, T.L. (1986). Teacher behavior and student achievement. In M.C. Wittrock (Ed.)., Handbook of Research on Teaching (pp. 328-375, 3rd ed.). New York: MacMillan.

Bruner, J.S. (1963). *The process of education*. Cambridge: Harvard University Press.

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

Cocking, R.R. & Renninger, K.A. (Eds.), 1993. *The development and meaning of psychological distance*. Hillsdale, New Jersey: Lawrence Erlbaum Associates.

Creemers (1992), *Effectieve instructie: een empirische bijdrage aan de verbetering van het onderwijs in de klas*, 's-Gravenhage: SVO.

Crooks, T. J. (1988). The impact of classroom evaluation practices on students. *Review of Educational Research,* 58, 438-481.

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

Darling-Hammond, L. (1989). Accountability for professional practice. *Teachers College Record*, *91*(1), 59-80.

Dochy, F.J.R.C., Moerkerke, G., & Martens, R. (1996). Integrating assessment, learning and instruction: Asessment of domain-specific and domain-traxlscending prior knowledge and progress. *Studies in Educational Evaluation, 22,* 309-339.

Doyle, W. (1983). Academic work. *Review of Educational Research*, *53*, 159-199.

Evertson, C.E., & Green, J.L. (1986). Observation as inquiry method. In M. C. Wittrock (Ed.), *Handbook of research on teaching.* (pp. 162-213). New York: MacMillan.

Fredericksen, J. R., & Collins, A. (1989). A systems approach to educational testing. *Educational Researcher,* 18(9), 27-32.

Frederickson, N. (1984). The real test bias: Influences of testing on teaching 695-699. 438-481. and learning. *American Psychologist*

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

Getzels J.W. & Jackson, P.W. (1963). The teacher's personality and characteristics. In N.L. Gage (ed.). *Handbook of research on education*. Chicago: Rand McNally.

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

Haertel, E.H. (1991). New forms of teacher assessment. In G.Grant (ed.), *Review of Research in Education.* (pp. 3-29). Washington: American Educational Research Association.

Jaeger, R.M.(1993). *Live vs. Memorex: psychometric and practical issues in the collection of data on teachers' performances in the classroom*. Center for Educational Research and Evaluation. University of North Carolina. (1993). ERIC reproduction services no: ED360325

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

Kagan, D.M. (1990). Ways of evaluating teacher cognition: inferences concerning the Goldilocks principle. *Review of Educational Research*, *60*(3), 419-469.

Kane, Crooks & Cohen, (1999). Validating Measures of Performance. *Educational Measurement: Issues and Practice*. *Educational Measurement: Issues and Practice, 18(2),* 5-17.

Kane, M.T. (2004). Certification Testing as an Illustration of Argument-Based Validation. *Measurement, 2* (3), 135-170.

Levine, M. (Ed.). (1988). *Professional practice schools: Building a model*. Washington: American Federation of Teachers. ED 313 344

Madaus, G. F. (1988). The influences of testing on the curriculum. In L. N. Tarner (Ed.), *Critical issues in curriculum* (pp. 83-121). Eighty-Seventh Yearbook of the National Society for the Study of Education, Part I. Chicago: University of Chicago Press.

Mehrens, W. A. (1992). Using performance assessment for accountability purposes. *Educational Measurement: Issues and Practice,* 11(1)3, -9, 20.

Messick, S. (1994). The interplay of evidence and consequences in the validation of performance assessments. *Educational Researcher,* 23(2), 13-23.

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

Nägele, R.C & Vissers, J.A.M.M. (2003) *Rijopleiding in Stappen (RIS). Evaluatie van de vervolgproef in de provincie Gelderland 2002-2003. 'Driver Training Stepwise (DTS). [Evaluation of the follow-up in the province of Gelderland].* Veenendaal, Traffic Test.

Prodromou, L. (1995). The backwash effect: From testing to teaching. *EL T Journal, 49,* 13-25.

Reeves, T. C. (2006). Design research from the technology perspective. In J. V. Akker, K. Gravemeijer, S. McKenney, & N. Nieveen (Eds.), *Educational design research* (pp. 86-109). London: Routledge.

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

Sherin, M. (2007). The development of teachers' professional vision in video clubs. In R. Goldman & R. Pea & B. Barron & S. Darry (Eds.), *Video research in the learning sciences* (pp. 383-395). Mahwah, NJ: Lawrence Erlbaum Associates.

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

Simon, A., & Boyer, E.G. (1974). *Mirrors for behavior. An anthology of classroom observation instruments*. Wyncote: communication Material Center.

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

Stodolsky, S.S. (2002). Classroom observation. In Millmann, J., Darling-Hammond, & L. (Eds.), *The new Handbook of Teacher Evaluation. Assessing elementary and secondary school teachers.* (pp. 175-190). Newbury Park, California: Corwin Press.

Swanson, D., Norman, G., & Linn, R. L. (1995). Performance-based assessment: Lessons from the health professions. *Educational Researcher,* 24(5), 5-11, 35.

Tom, A.R., & Valli, L. (1990). Professional knowledge for teachers. In R.W. Houston (Ed.), *Handbook of research on teacher education* (pp. 372-392). New York: MacMillan.

Van den Akker, J. (1999). Principles and methods of development research. In J. van den Akker, N. Nieveen, R. M. Branch, K. L. Gustafson & T. Plomp (Eds.), *Design methodology and developmental research in education and training* (pp. 1-14). The Netherlands: Kluwer Academic Publishers.

van Es, E. A., & Sherin, M. G. (2006). How different video club designs support teachers in "learning to notice". *Journal of Computing in Teacher Education, 22*(4), 125-135.

Van Kuyk, J. J. (2000). *Piramide. Educatieve methode voor drie-zesjarige kinderen*. *Wetenschappelijke verantwoording.* Arnhem: Citogroep.

Veenman, S. (1984). Perceived problems of beginning teachers. *Review of Educational Research*, *54*(2). 143-178.

Verloop, N. (1988). Investigating teacher cognitions. *Journal of Curriculum Studies*, *20*, 81-86.

Vermunt, J. & Verschaffel. (2000). Process oriented teaching. In P.R.J. Simons, J. van der Linden & T. Duffy. *New Learning* (209-225). Dordrecht: Kluwer Academic Publishers.

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

Vu, N., & Barrows, H. (1994). Use of standardized patients in clinical assessments: Recent developments and measurement findings. *Educational Researcher,* 23(3), 23-30.

Wiggins, G. (1989). A true test: Toward more authentic and equitable assessment. *Phi Delta Kappan,* 70,703-713.

Wiggins, G. (1993). *Assessing student performance: Exploring the purpose and limits of testing.* San Francisco: Jossey-Bass.

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.

Table 3: Perceived quality of evidence as evaluated by teacher consultants. Context 1: instruction in Kindergarten for concept acquisition.

	Characteristics video footage		Perceptibility video footage					Characteristics supporting evidence						Asessability ofVP	
Loc-ation	**Overall picture**	**Camera position**	**Classroom Setting**	**Teacher visible and audible**	**Student visible and audible**	**Blackboard readable?**	**Other learning media visible**	**Lesson plan Available ?**	**Theme**	**Subject**	**Context description available**	**Learning materials added?**	**Decision process reported?**	**Process Relevance all evidence**	**Representativeness of the set of situations**
W	Teacher and students	Two hand held cameras	Small groups	+	+	NA	+	Project manual: detailed plan	Autumn	Integrated subjects		Yes	Yes	+ Instruction for concept acquisition	+_
E	Teacher in group	Two hand held cameras	Whole group	++	+-	NA	+-	Project manual: detailed plan	Moving into another house	Integrated Subjects		Yes	Yes	+ Instruction for concept acquisition	+-
L	Teacher in group	Two hand held cameras	Whole group	++	+-	NA	+-	Project manual: detailed plan	Saint Nicholas	Integrated Subjects		Yes	Yes	+ Instruction for concept acquisition	+-

NA= not applicable; - = weak quality, +- = questionable quality, + = good quality ? = not available

Table 4: Perceived quality of evidence as evaluated by teacher consultants. Context 2: coaching in senior vocational education

	Characteristics video footage		**Perceptibility video footage**					**Characteristics supporting evidence**						**Asessability ofVP**	
Location	**Overall picture**	**Camera position**	**Classroom Setting**	**Teacher visible and audible**	**Student visible and audible**	**Blackboard readable?**	**Other learning media visible**	**Lesson plan Available ?**	**Theme**	**Subject**	**Context description available**	**Learning materials added?**	**Decision process reported?**	**Process Relevance of evidence**	**Representativeness of the set of situations**
A	Teacher and students	Three manually operated fixed cameras	Small group (2 students)	+	+	NA	+	Yes	Technical design	Technology of infrastructure	Yes	Yes	Yes	+	+
Ha	Teacher and students	Three manually operated fixed cameras	Small group (3 students)	+	+	NA	+	Yes	Technical design	Technology of infrastructure	Yes	Yes	Yes	+	+
Hb	Teacher and students	Three manually operated fixed cameras	Small group (4 students)	+	+	NA	+	Yes	Technical design	Architecture	Yes	Yes	Yes	+	+
R	Teacher and students	Three manually operated fixed cameras	Small groups (4 students)	+	+	NA	+	Yes	Technical design	Architecture	Yes	Yes	Yes	+	+

NA= not applicable; - = weak quality, +- = questionable quality, + = good quality ? = not available

Table 5: condensed observations about video episodes within portfolios (Driver training).

Location	Teacher visible and audible	Overall picture	Student visible and audible	Blackboard readable?	Other learning media visible	Camera position	Lesson plan Available ?	Theme	Subject	Setting	Relevance of situation	Context description available	Learning materials added?
All locations: 4 instructors	I: NA + S: NA + C: NA +	Front screen Rear mirror	+ +	NA	+	Front screen Rear mirror	- - -	Variable Variable Variable	Vehicle control Simple situations Complex situations	Individual tutoring Individual tutoring Individual tutoring	Instruction and coaching	+	-

NA= not applicable; - = weak quality, +- = questionable quality, + = good quality ? = not available

Table 6: Elaboration of assessment tasks for students as a basis of their video portfolios

Location	**Assessment task**	**Instructions and hints related to collection of video evidence**	**Technical instructions**
E	Uncover and picture students' misconceptions during own designed learning activities, fostering a dialogue between students about a physics theme. Parts: 1. Episode showing the students displaying their misconception. 2. Episode showing a discussion of the first episode via stimulated recall (with colleague) 3. Episode showing a discussion with the students in uncovering and correcting the misconception.	Yes, Literal instruction: Devise an experiment in such a way that it enables students to move and grab objects to make their conceptions clear. - Look for a quiet place with favorable light(no from light and sound conditions) - The students are taken out of their classrooms. - Choose intermediate level students or a little bit weaker. - Make boys and girls only dyds of equal level, to ensure equal participation in tasks.	Yes, via special manuals explaining the editing and uploading of video episodes.
G	Deliver an instruction for an assignment to be carried out independently. The instruction will be specific and clear. The basis of the student assignment is a historical problem or theme. a. Upload a video episode in which you show a clear instruction for the assignment. Describe why you chose this particular episode. b. Upload one or more video episodes in which the activities around the assignment are ended well, that is in a recognizable way.	No	Unknown
A1	In this VP you image your development of competence in carrying out lessons. You will make recording during different moments during your teacher training. Next, you will choose the episodes which show the level of competence of that moment. The video episode, completed with other documents are placed in a story line. Two assignments: -Communication, leadership and classroom management -Coaching of learning processes. 1: Pay attention to the development of your repertoire regarding: - starting, rounding up, and transitions in a lesson;- complete task instruction; - use of questions, warranting individual accountability of students ; motivating classes and students; guidance of individuals or groups of students 2: Pay attention to: - making contact with students; - create safe environment; leadership; give responsibility to students; maintaining group discipline Give a title to your story line. Choose (1) an episode showing your entry level. (2) After the course about communication and learning processes (3) at the end of your teacher training. Explain why you chose these episodes. Explain which entries of the assessment rubrics can be illustrated. What is your level and how does the episode show this? Which moments make you feel proud? What would you like to develop further?	Within teacher training some attention is paid to the choice of teaching episodes.	Yes, wa part of four training sessions: making video registrations, editing, export into DIVI-dossier format, upload and integrate into DIVI-dossier environment.
A2	1. Self portrait: portrait yourself using a film of maximum three minutes. 2. Design two lessons, of which you choose one to carry out. Show by means of maximum 4 video episodes: a. for science and math teachers: - how you deliver direct instruction;- how you use concrete material in your lesson (practicum, demonstration) b. for language teachers: - that you use authentic reading or listening texts and images in your lesson. choose two of the three models of teaching and show how: - you deliver direct instruction; and/or -how you lead a Socratic conversation; and/or you use one of the basics structures of cooperative learning; c. all teachers - how you interact with students; - how your students have actually gained knowledge. 3. Story line: Add to [your sources] a filled out competence rubric. You create a story line of one of the domains of competence. Components: video episodes with commenting texts between the episodes. Complete the picture with additional documents from [My sources].	No	via special manuals explaining the editing and uploading of video episodes.

Table 7: characteristics and observations regarding video portfolios

Location and learner teacher	Overall picture	Camera position	Classroom setting	Teacher Visible and audible	Student visible and audible	Blackboard readable?	Other learning media visible	Lesson plan Available ?	Theme	Subject	Relevance of situation	Context description available	Decision process reported?	Learning materials added?
E1	a. Conversation between students	a. Fixed on two students	a. Dyads	+ +	+	NA	No	No	Energy (throwing a ball)	Physics	Coaching of learning processes	No	No	No
	b. teacher and two students	b. Fixed on teacher and 2 students	b. Teacher and dyad	+ +	+	-	No	No						
E2	a. Conversation between students	a. Fixed on two students	a. Dyads	+ +	+	NA	No	No	Gravitation (bowling ball, soccer ball)	Physics	Coaching of learning processes	No	No	No
	b. teacher and two students	b. Fixed on teacher and 2 students	b. Teacher and dyad	+ +	+	-	No	No						
E1	a. Conversation between students	a. Fixed on two students	a. Dyads	+ +	+	NA	No	No	Gravitation (parachute dive)	Physics	Coaching of learning processes	No	No	No
	b. teacher and two students	b. Fixed on teacher and 2 students	b. Teacher and dyad	+ +	+	-	No	No						
G1	Directed at listening students. Camera off scene and shaky zooms	Fixed	Groups within whole class	+ +	- -	NA	No	Yes	Violence	History	Instruction	No	Yes	Assignment
G2	Only teacher	Fixed	Whole class	+ +	- -	-	No	Yes	?	History	None	No	Yes	Powerpoint
G3	Teacher and part of students	Fixed Camera man	Whole class	+ +	- - .	-	No	Yes	Developmental help	Geography	Instruction	Yes	Yes	Powerpoint
A1	Moving camera; directed at one student	Cameraman: hand-held	Whole class	- -	-	NA	-	-	Off task student	Geography	None	No	Yes	No
A2	Fixed on one part of group	Fixed in back of room; wide angle lens	Whole class: seat work	- -	- camera off spot	NA	-	-	Off task student	Biology	Classroom management	No	Yes	No
A3	only teacher in front of class (shaky picture)	Fixed side of class	Whole class	+- +	-	-	-	-	Discussing home work	Chemistry	None	No	Yes	No
J1	a. Teacher and part of students	a. Camera operator	a. Small plenary group	+ +	+ +	-	Yes	Yes	Spanish lexicon 'Tapas'	Spanish	Instruction	Yes	Yes	No
	b. Teacher front class	b. Fixed	b. Small plenary group	+ +	+ +	NA	Yes	Yes	Prepare meal		Coaching of learning processes	Yes	Yes	No
J2	Teacher front class	Camera man: hand held.	Large plenary group	+ +	- -	+	No	Yes	Grammar	French	Interpersonal competence	Partly	Yes	No

NA= not applicable; - = weak quality, +- = questionable quality, + = good quality ? = not available