EDUCATIVE ASSESSMENT FOR/OF TEACHER COMPETENCY

Malmö Studies in Educational Sciences No. 41
Studies in Science and Technology Education No. 18

ISBN 978-91-977100-3-9
ISSN 1651-4513
ISSN 1652-5051
Holmbergs, Malmö 2008

ANDERS JÖNSSON

EDUCATIVE ASSESSMENT FOR/OF TEACHER COMPETENCY

A study of assessment and learning in the "Interactive examination" for student teachers

Malmö University, 2008
School of Teacher Education

The publication will also be made available electronically,
see www.mah.se/muep

TABLE OF CONTENTS

ACKNOWLEDGEMENTS 11
ABSTRACT 13
PAPERS INCLUDED IN THE DISSERTATION 15
LIST OF FIGURES 17
LIST OF TABLES 19
PREFACE 21
Setting the scene 21
Reading instructions 22
INTRODUCTION 25
Performance assessment versus testing 26
Authentic assessment 29
Problems of introducing authentic assessment 31
The problem of credibility 32
Reliability issues 32
Validity issues 34
The problem of credibility: Conclusions 41
The question of student learning 42
Feedback 47
Self-assessment 48
Multiple levels of success 50
The question of student learning: Conclusions 50
STUDY I: THE USE OF RUBRICS 53
Research questions 55
Does the use of rubrics enhance the reliability of scoring? 56
Can rubrics facilitate valid judgment of performance assessments? 58

Does the use of rubrics promote learning and/or improve instruction? ... 59
Perceptions of using rubrics ... 60
Interpretation of criteria ... 60
Student improvement ... 61
AUTHENTIC ASSESSMENT: PUTTING IT INTO PRACTICE ... 65
Context ... 67
Criteria for teacher competency ... 69
To articulate "tacit knowledge" ... 71
Formulating criteria ... 72
The "Interactive Examination" for dental students ... 73
The "Interactive examination" for student teachers ... 76
The personal task ... 76
The professional document ... 79
The rubric ... 79
Comparison of quantitative self-assessment ... 81
Developmental changes in the "Interactive examination" for student teachers ... 82
Research methodology ... 84
Research questions ... 84
Sample ... 85
Research data and analyses ... 85
Methodological limitations ... 87
STUDY II-IV: THE "INTERACTIVE EXAMINATION" ... 91
Study II: Does the "Interactive examination" for student teachers work? ... 91
Results and conclusions ... 91
Study III: Is the "Interactive examination" for student teachers valid for its summative and formative purposes? ... 93
Results and conclusions ... 96
Study IV: Does the use of transparency improve student performance? ... 97
Results and conclusions ... 98
DISCUSSION ... 99
Assessing teacher competency ... 100
Authenticity of the "Interactive examination" ... 103
A systems approach to assessment ... 105
Assessing teacher competency: Conclusions ... 106

Assessing self-assessment skills 107
The professional document 108
Assessing self-assessment skills: Conclusions 109
Supporting student performance 109
Supporting student performance: Conclusions 111
Unique features in the "Interactive examination" 112
Self-assessment 112
The scoring rubric 113
Transparency 115
The use of information- and communication technology 116
Unique features: Conclusions 117
Contributions to research 118
Future research 118
Extrapolation to workplace settings 119
Effects on student motivation and learning, and on teachers' work 120
Implications for practice 120
Implications for the design of performance assessments 120
Implications for teacher education 122
REFERENCES 123
APPENDICES 135
Appendix A. The "Interactive examination" 135
Appendix B. Scoring rubric for the "Interactive examination" 142
Appendix C. Excerpts from the exemplars 146
Appendix D. References to papers from the Xpand project 148
PAPERS I-IV (PUBLISHED VERSION ONLY) 151

ACKNOWLEDGEMENTS

Some words of special thanks to those who contributed to the development and quality of this dissertation: My supervisor (Gunilla Svingby); the members of the Xpand research group at Malmö University; the ”assessment people” at Stockholm University (Lars Lindström, Viveca Lindberg, Astrid Pettersson, Lisa Björklund Boistrup, Helena Tsagalidis, and others), the discussants of the not-yet-finished versions of my manuscript (Lars Lindström, Jan-Eric Gustafsson, Ulla Tebelius), and those checking the manuscript for the final seminar (Sven Persson, Margareta Ekborg, Harriet Axelsson). Also, a very special thanks to Sven-Åke Lennung, for all his support and his interest in my work.

Finally, it should be acknowledged that the development of the ”Interactive examination” was funded by the former national agency for distance education (DISTUM).

ABSTRACT

The aim of this dissertation is to explore some of the problems associated with introducing authentic assessment in teacher education. In the first part of the dissertation the question is investigated, through a literature review, whether the use of scoring rubrics can aid in supporting credible assessment of complex performance, and at the same time support student learning of such complex performance. In the second part, the conclusions arrived at from the first part are implemented into the design of the so-called "Interactive examination" for student teachers, which is designed to be an authentic assessment for teacher competency. In this examination, the students are shown short video sequences displaying critical classroom situations, and are then asked to describe, analyze, and suggest ways to handle the situations, as well as reflect on their own answers. It is investigated whether the competencies aimed for in the "Interactive examination" can be assessed in a credible manner, and whether the examination methodology supports student learning. From these investigations, involving three consecutive cohorts of student teachers (n = 462), it is argued that three main contributions to research have been made. First, by reviewing empirical research on performance assessment and scoring rubrics, a set of assumptions has been reached on how to design authentic assessments that both support student learning, and provide reliable and valid data on student performance. Second, by articulating teacher competency in the form of criteria and standards, it is possible to assess students' skills in analyzing classroom situations, as well as their self-assessment skills. Furthermore, it is demonstrated that by making the assessment demands transparent, students' per-

formances are greatly improved. Third, it is shown how teacher competency can be assessed in a valid way, without compromising the reliability. Thus the dissertation gives an illustration of how formative and summative purposes might co-exist within the boundaries of the same (educative) assessment.

Keywords: authentic assessment, formative assessment, learning, reliability, performance assessment, scoring rubrics, teacher education, validity

PAPERS INCLUDED IN THE DISSERTATION

Paper I

The use of scoring rubrics: Reliability, validity and educational consequences
Co-author: Svingby, Gunilla
Published: 2007, Educational Research Review, Vol. 2, pp. 130-144.

Paper II

Dynamic assessment and the "Interactive examination"
Co-authors: Mattheos, Nikos; Svingby, Gunilla, & Attström, Rolf
Published: 2007, Educational Technology & Society, Vol. 10, pp. 17-27.
Presented: EARLI Conference, Nicosia, Cyprus, August 2005.

Paper III

Estimating the quality of performance assessments: The case of an "Interactive examination" for teacher competency
Co-authors: Baartman, Liesbeth & Lennung, Sven A.
Manuscript submitted for publication in Learning Environments Research.
Presented: EARLI Conference, Budapest, Hungary, September 2007.

Paper IV

The use of transparency in the "Interactive examination" for student teachers

Manuscript submitted for publication in Assessment in Education: Principles, Policy & Practice.

Presented: AEA Europe Conference, Stockholm, Sweden, November 2007.

LIST OF FIGURES

Figure	Heading	Page
1	A simplified example of a scoring rubric	54
2	A graphic representation of the six stages in the "Interactive examination"	74
3	The "Wheel of competency assessment"	94

LIST OF TABLES

Table	Heading	Page
1	Examples of how course objectives were operationalized in the rubric	81
2	An overview of data collected, and analyses performed, in relation to the different studies on the "Interactive examination"	86

PREFACE

Setting the scene

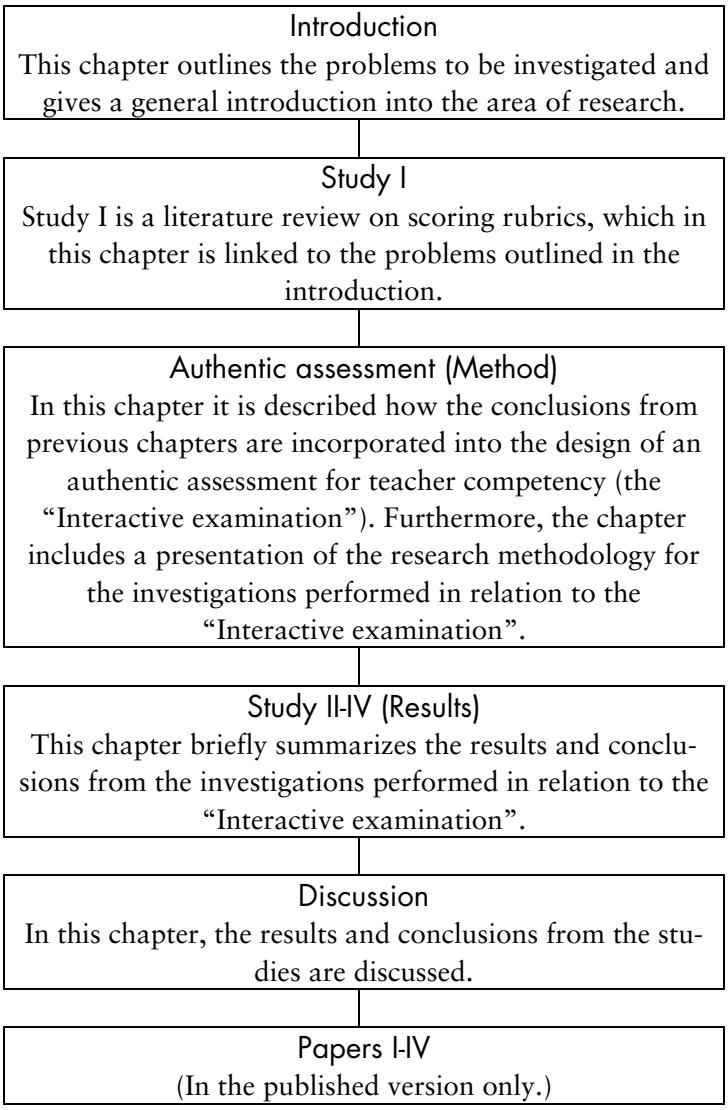
The research presented in this dissertation is part of a larger project (the "Xpand" project), involving all students in the teacher-education program at Malmö University, with Science, Mathematics, or Geography as their subject major, during their first semester. The core assumption in this project is that individuals' capability to identify their own actual competency, and realize when actual competency differs from intended (i.e. professional) competency, is central to competency development. The capacity to understand and articulate one's own competency, and to identify alternatives, thus makes development possible.

Besides the "Interactive examination", which can be described as an authentic assessment for teacher competency and which is the focus of this dissertation, other tools and applications have also been developed and evaluated within the project. These tools, which are thought to support student reflection through various forms of self-assessment, are of two kinds. One tool involves the individual student in self-reflection through self-reported Likert scales (e.g. epistemological beliefs, academic confidence, etc.). The other tool focuses on group dynamics. The ability to work effectively in teams or groups is often taken for granted, in spite of frequent experiences of conflicts. However, professional competency of teachers includes working in groups, and in order to help developing such competency the tool allows for analysis of group dynamics and of the quality of net-based dialogues. Through a combination of net-based utilities (such as Social Network Analysis and

the labeling of own contributions in discussion fora) the group or the individual (or an educator) can easily analyze group processes and the specific contributions of individual members of the group. For more thorough descriptions of the tools and the research performed, see references in Appendix D.

Reading instructions

This dissertation consists of four papers, together with an "extended summary", including problem statement, a general introduction into the area of research, a chapter dealing with methodological issues, brief descriptions of the results, and an overarching discussion (see overview on the next page). Following the Swedish tradition, the papers are attached at the end (Note: The papers are attached in the published version only), and not as chapters in the dissertation. Efforts have been made, however, to make it possible to read the summary from the beginning to the end without necessarily making constant leaps back and forth between summary and papers. This means that some information is found both in the summary and in the papers. Detailed information on analyses made and specific results are, however, confined to the papers.



Schematic overview of the chapters in the dissertation.

INTRODUCTION

Teacher education is a profession-directed education, aiming for students to become competent professionals. Aiming for competency means that the students have to develop their knowledge, skills, and attitudes into integrated and situation-relevant actions, in order to master relevant tasks (Taconis, Van der Plas, & Van der Sanden, 2004). To be "competent" thus means to be able to *act knowledgeably* in relevant situations.

Aiming for competency also means that there is a need for assessment methodologies which *assess the acquisition* of such competencies. Since most summative assessments give consequences in terms of grades or certification (i.e. they are "high-stakes"), such assessments have been shown to steer student learning (e.g. Struyven, Dochy, & Janssens, 2005). This effect, which is sometimes unintentional, is often called the "backwash" of assessment (Biggs, 1999). However, if summative assessments were designed so that they could be used for formative purposes as well, they would not have to be limited to only measuring students' acquisition of the competencies aimed for, but could also be used to *support the development* of the same competencies (Baartman, Bastiaens, Kirschner, & Van der Vleuten, 2008; Black, Harrison, Lee, Marshall, & Wiliam, 2003). It is argued that such combinations of formative and summative assessment are imperative in educational settings, due to the strong effect of assessment on student learning, and this argument will permeate the work in this dissertation:

> You can't beat backwash, so join it. Students will always second guess the assessment task, and then learn what they think will meet those requirements. But if those assessment requirements mirror the curriculum, there is no problem. Students will be learning what they are supposed to be learning. (Biggs, 1999, p. 35)

Another question of particular interest when educating professionals, like teachers, is how the students can be prepared for lifelong learning and provided with a continuing ambition to improve their work (Hammerness et al., 2005) – an often stated aim of higher education in general (e.g. Birenbaum, 2003; Segers, Dochy, & Cascallar, 2003). One answer to this question is that teacher education must supply the students with the necessary skills for self-assessing their own performance as teachers and to change it, if required. However (again due to the strong effect of assessment on student learning), students' skills in reflecting on their performance must not only be taught, but also be assessed. An assessment methodology which could assess students' self-assessment skills, and in this way help them in developing these skills, would therefore make a substantive contribution to teacher education.

In line with the arguments above, the aim of this dissertation is to explore how teacher competency (including self-assessment skills) can be assessed in an authentic manner, and how the assessment can support student learning, while still acknowledging the importance of credibility and trustworthiness in the assessment (i.e. "educative assessment").

Performance assessment versus testing

When assessing competency, it could be argued that if we want to know how well somebody can perform a certain task, the most natural thing would be to ask her to do it, and then assess her performance (Kane, Crooks, & Cohen, 1999). Such assessments, where students are assessed during actual performance, are called "performance assessments".

Performance assessments are characterized by two things. First, the students are assessed while actually performing, which means that the assessment is "direct", and that inferences to theoretical constructs (like "understanding" or "intelligence") do not have to be made. Second, performance-assessment tasks can be positioned in the far end of the continuum representing allowed openness of student responses as opposed to multiple-choice assessments (Messick, 1996). Such open-ended tasks are needed if complex competencies are to be assessed.

The importance of introducing performance assessment when assessing competency is best seen in the light of current assessment practices in higher education. Although these assessment practices may vary between different countries and between different subjects, they often share some common characteristics. For instance, as a student in higher education you are likely to encounter written exams, or tests. During such a test you are required to give the correct answers to a number of questions during a specified length of time. Furthermore, the test is typically taken by individuals in isolation, which means that neither tools nor collaboration is allowed. Even though this kind of assessment practice is very common, it has been criticized for being summative, decontextualized, and inauthentic (e.g. Birenbaum et al., 2006).

That a test is *summative* means that it is primarily designed to measure students' knowledge, not to improve it. A summative test does not aim at providing any feedback to the student regarding her specific strengths or weaknesses, or her progress. Instead, the feedback is often restricted to an overall score, a grade, or in the case of norm-referenced assessment, a rank in relation to other students. Lacking adequate feedback, summative tests fail to support and encourage relevant student learning.

That a test is *decontextualized* means that the items are not tied to any particular situation. Instead, the knowledge measured is thought to be generic and applicable in many different contexts. This, however, is not in line with the assumption that human knowledge is highly contextualized (Biggs, 1996; Shepard, 2002; Wertsch, 1991, 1998), and it has been argued that students (when lacking a given context) have to apply an artificial and very test-specific context (Spurling, 1979). This implies that the knowledge

measured is not generic and applicable in many different contexts, but, quite on the contrary, closely tied to the test situation.

Furthermore, since decontextualized tests are not supposed to measure students' knowledge in context, inferences have to be made from student performance on the assessment to an underlying theoretical construct (Frederiksen & Collins, 1989; Kane et al., 1999). This "indirect" way of testing, just like the summative feature, makes it difficult to support and encourage student learning, since test outcomes in terms of "understanding", "ability", or "achievement" are difficult to relate to actual performance.

Another feature of decontextualized tests is that student performance is typically broken down to discrete items or well defined problems. Such a fragmentation has been argued to reward primarily atomized knowledge and rote learning, rather than complex and authentic competencies (Birenbaum et al., 2006; Gipps, 2001; Shepard, 2000). Since assessment strongly affects student learning (e.g. Struyven et al., 2005), tests that focus on recall might (unintentionally) steer student learning towards surface approaches to learning (i.e. through the "backwash" of assessment).

That a test is *inauthentic* means that the students are not assessed as to whether they can (or cannot) do the things they are intended to do in "real life" or in professional settings. Instead, students are assessed with an instrument (a written test) and in a specific situation (e.g. individually and with no tools) which does not resemble the authentic context in which the students are supposed to use their knowledge. Thus inferences about students' knowledge are made from performances of a different kind than the actual expected performance. This problem is closely related to the issues on decontextualized tests discussed above. However, the notion of "authenticity" implies that the assessment should not be tied to *any* given context (such as being limited to a school setting or an imaginary context). Rather, the assessment should replicate the circumstances of the specific "communities of practice" which the students are to become participants of, so that the students can strive for what is considered excellent performance within these communities (Lave & Wenger, 1991; Wiggins, 1998). For academically-oriented education, such communities of practice could be the liberal arts (viz the social and natural sciences, fine arts, litera-

ture, and the humanities), and for profession-directed education it could be the professional institutions (e.g. schools, hospitals, or law offices).

In summary, the criticism against summative, decontextualized, and inauthentic tests points to some quite severe problems, namely that:

- such tests do not support relevant student learning,
- the knowledge assessed might be limited to the test situation,
- such tests steer students' learning towards atomized knowledge and rote learning, and
- inferences about student knowledge are made from performances of a different kind than the performance educated for.

Authentic assessment

As opposed to summative, decontextualized, and inauthentic tests, performance assessment deals with "activities which can be direct models of the reality" (Black, 1998, p. 87). However, since an assessment can be both "direct" and open-ended, without having any connection to an authentic context, some authors instead prefer to use the concept "authentic assessment", which denotes that:

1. The assessment tries to reflect the complexity of the real world and provides more valid data about student competency, by letting the students solve realistic problems (Darling-Hammond & Snyder, 2000).
2. Assessment criteria, as well as standards for excellent performance, reflect what is considered quality within a specified community of practice (Wiggins, 1998).

The strength of authentic assessment is that inferences about student competency are made from performances of a kind *similar* to the performance educated for. There can never be a perfect match between assessments and "reality", however, since restrictions of some kind are always imposed for practical and logistical reasons,

making all assessments artificial in some way (Kane et al., 1999)[1]. Still, by designing the assessment with as many authentic dimensions as possible (e.g. the task, the social or physical context, etc.) it has been argued that authentic assessment can provide more valid data about student competency as well as having a positive impact on student learning. The latter assumption rests on the "backwash effect" (i.e. that assessment affects student learning) referred to previously, when students are supposed to learn complex competencies (Gulikers, Bastiaens, & Kirschner, 2004). Furthermore, it could be assumed that authenticity facilitates transfer to the target domain[2] (Havnes, 2008).

In summary, introducing authentic assessment in teacher education assumes that:

- student learning is directed towards those complex competencies assessed – as opposed to favoring atomized knowledge and rote learning,
- inferences about student competencies are more valid than when made from performances of a different kind than the performance educated for, and
- the competencies assessed are not limited to the test situation.

[1] A "perfect match" between assessment and real-life settings is not even necessarily desirable, since there might be aspects of the educational context (such as a greater tolerance for failure) that are needed in order to foster thoughtful learning (Lindström, 2008).

[2] A distinction between "domain" and "construct" is made in this dissertation. "Domain" refers to the "community of practice" (see Lave & Wenger, 1991; Wenger, 1998) to which the assessment performance is thought to generalize or extrapolate. "Construct", on the other hand, is used to denote theoretical constructs, such as "intelligence" or "understanding". However, terms like "construct-irrelevant difficulty" have not been changed when referring to the wording of specific authors (like Messick, 1996).

Problems of introducing authentic assessment

Before introducing authentic assessment in educational settings, there are some difficulties that need to be considered:

1. The problem of whether assessment of complex performance can be carried out in a credible and trustworthy manner.
2. The question whether assessing performance actually supports student learning of complex competencies.
3. It could be assumed that performance assessments are more time consuming and costly than paper-and-pencil tests.

All of these issues are of great significance. Since the last one can (at least to some extent) be overcome by the use of modern information- and communication technology (ICT), we will return to this issue later, and the first two problems will constitute the main foci of interest in this initial part of the dissertation.

The importance of credibility and trustworthiness becomes clear if we were to suppose that assessments were *not* credible and trustworthy. For example, it would then be left to chance to decide whether the students succeed or not. Since higher education is often a high-stakes enterprise for the students, affecting their future to a large extent, this is not satisfactory. Furthermore, if the assessments of teacher performance were not credible and trustworthy, this could mean that the wrong kind of performance was rewarded. As a consequence, students who are in fact good teachers would not necessarily be the ones who received high grades. But even disregarding issues of grading and other high-stakes decisions, assessments that are not credible and trustworthy would also fail to direct student learning. This is because such assessments do not provide systematic and consistent feedback (due to the large influence of chance and/or subjectivity); and by potentially rewarding the wrong kind of performance, student learning would be misguided.

The question whether assessing performance actually supports student learning of complex competencies is also of great importance, since if it does not, the incentive to introduce performance assessments in higher education would be greatly weakened.

The remaining sections in this introductory chapter attempts to give a more detailed picture of the problems of introducing authentic assessment in higher education in general, in relation to the issues of student learning and credible assessment. The chapter then concludes with a set of empirically grounded assumptions on how to deal with these problems.

The problem of credibility

The problem of assessing complex performance in a credible way is often argued to be most pressing for high-stakes summative assessments. This is because these assessments have serious consequences for those being assessed, in particular concerning what kind of education and job they will have access to. Institutions using performance assessment for high-stake decisions are thus faced with the challenge of showing that evidence derived from these assessments is both valid and reliable. Classroom assessments, on the other hand, are often seen to be less in need of high levels of reliability, since decisions made on the basis of classroom assessment can easily be changed if they turn out to be wrong (Black, 1998). Still, as was argued previously, if assessments are to have the potential to direct student learning by providing systematic and consistent feedback, and by rewarding the appropriate performance, all assessments need to be both valid and reliable (even if lower levels of reliability might be considered acceptable for classroom assessments).

Reliability issues

Assessments have to be evidence-based and performed with disinterested judgment, but the interpretation of both evidence and judgment has to be made by some individual, and there is always the question whether another person would come to the same conclusion. Ideally, an assessment should be independent of who does the scoring, and the results should be similar no matter when and where the assessment is carried out, even if this is hardly obtainable. This goal can be reached to a greater or lesser extent, however,

and the more consistent the scores are, the more reliable the assessment is thought to be (Moskal & Leydens, 2000).

What factors then, might threaten reliability? Dunbar, Koretz, and Hoover (1991) show, through six different studies investigating the reliability of writing performance, that *assessor reliability* can vary considerably depending on the number of points on the scoring scale and on the conditions of the assessment (for example natural settings versus controlled experimental conditions). Their study also shows that assessor reliability often is quite low, at least when compared to the standards of "traditional testing" (i.e. tests within the psychometric tradition). As Brennan (2000) notes, low levels of reliability typically occur when students choose their own tasks or produce unique items, while on the other hand inter-rater reliability tends to be high when tasks are standardized. One major reason for the low reliability of performance assessments is therefore likely to derive from the fact that they are open-ended, and as a way to remedy the situation of low reliability, restrictions could be added to the assessment. But if severe restrictions are imposed, due to calls for high reliability, does the assessment still "capture" the full scope of what it was intended to "capture"? This is a classical reliability versus validity dilemma, where low reliability can be raised by defining the task more strictly, but this would at the same time affect validity negatively (Brennan, op. cit.; Dunbar et al., op. cit.). When using authentic assessment, this tradeoff – where the validity of the assessment is "sacrificed" to obtain higher levels of reliability – would not be acceptable. Instead, ways must be found of increasing the reliability, without losing track of what is considered important. Instrumental to increasing assessor reliability are detailed scoring protocols, sampled responses that exemplify the points on the scoring scale, and training of assessors (Dunbar et al., op. cit.; Linn, Baker, & Dunbar, 1994). Adding more assessors, however, does not increase the reliability in a significant way, and consequently there are often no reasons to employ more than one assessor (Brennan, op. cit.).

There are also other sources of error than the assessor. For example, the *variability of student performance on performance tasks* (even on tasks within the same domain) is often quite large, and this aspect might pose an even larger problem than assessor relia-

bility (Linn & Burton, 1994). By extrapolating results from the studies investigating reliability of writing performance, Dunbar et al. show that reliability increases markedly if more tasks are added to the assessment. For all of the studies except one, four tasks or less were enough to reach a reliability level of .7 or higher, which is generally considered sufficient (Brown, Glasswell, & Harland, 2004; Stemler, 2004). Similar results have been reported using generalizability theory (e.g. Baker, Abedi, Linn, & Niemi, 1995; Gao, Shavelson, & Baxter, 1994). Kane et al. (1999) thus argue that the assessment tasks should be relatively short, so that a number of different tasks can be used. Miller (1998), on the other hand, has shown that the type of task can influence the number of tasks required for acceptable levels of generalizability. By using longer and more complex tasks, fewer tasks are required to achieve satisfactory levels of generalizability. With complex, extended tasks, as few as two tasks could be sufficient, while shorter, open-ended tasks could require five to ten tasks.

Besides assessors and tasks, the reliability of performance assessments is affected by *variability due to occasions*. This aspect is included in the concept of intra-rater reliability, which measures the variability for the same assessor on more than one occasion. According to Brennan (2000), there are very few studies in the performance assessment literature that report on results from more than one occasion, but these findings suggest that this aspect might make a relatively small contribution as compared to the other sources of error.

To summarize, three main factors affect the reliability of performance assessments, namely assessors, tasks, and occasions. Of these, tasks and assessors are of primary importance, since they contribute most heavily to unwanted variability. To counterbalance the effect of assessors, detailed scoring protocols, sampled responses which exemplify the points on the scoring scale, and training of assessors have been suggested. In order to decrease variability due to tasks, more tasks could be added to the assessment.

Validity issues

When introducing performance assessment in higher education, one of the major threats to validity originates from the basic notion

that the tasks should be representative of the domain in question. Although this problem of "domain representation" might threaten the validity in all educational assessments, domain-irrelevant variance poses a greater threat for performance assessments. This is because performance assessments are typically open-ended and involve complex performance, thus possibly letting domain-irrelevant factors influence the assessment by being too broadly defined in relation to the domain (Mclellan, 2004; Messick, 1996).

There are two kinds of domain-irrelevant variance, which Messick calls "construct-irrelevant difficulty" and "construct-irrelevant easiness". "Construct-irrelevant difficulty" means that the task is more difficult for some individuals or groups, due to aspects of the task that are not part of the domain assessed. A classic example is the effect of reading comprehension when assessing subject-matter knowledge. "Construct-irrelevant easiness" on the other hand, could occur for example when the content of the task is highly familiar to some of the students. Whereas "construct-irrelevant difficulty" typically would lead to scores which are too low for the students affected, "construct-irrelevant easiness" would produce scores that are too high. In performance assessments, there are often contextual clues imbedded in the task, clues which help some students to perform appropriately. The context might also be more or less familiar to the students, making "construct-irrelevant easiness" a potential problem (Messick, 1996).

How then can these problems be approached? Basically, this depends on how validity is defined, and validity could either be seen as a property of the assessment, or as score interpretations and use (Black, 1998; Borsboom, Mellenbergh, & van Heerden, 2004). The first perspective is most widely used in natural sciences and psychological testing, whereas questions of validity in educational research are seldom confined to principles of measurement, but are rather seen to involve interpretations which stretch beyond the particular assessment. What needs to be valid, in such a perspective, is the *interpretation* of the scores, as well as the *use* of the assessment results (Borsboom et al., op. cit.; McMillan, 2004; Messick, 1996, 1998). However, the issue of in which ways these interpretations, and ways of using the assessment results, persist across different persons, groups, or contexts, is an empirical question. The valida-

tion process therefore becomes a matter of arguing from evidence supporting (or challenging) the intended *purpose* of the assessment. According to Messick, this view of validity integrates the forms of validity traditionally used into a unified framework of construct validity. In this framework, he distinguishes six aspects of construct validity, which may be discussed selectively, but none should be ignored. This means that when addressing the problem of validity in performance assessments, there is no neat little slice of validity (such as content validity) which can be cut off to be scrutinized, but rather that a comprehensive validation process must be performed, including both rational argument and empirical data, in order to claim validity of the assessment (Messick, op. cit.).

The six aspects of validity in Messick's framework are content, generalizability, external, structural, substantive, and consequential validity. Below, each aspect is described briefly, together with suggestions of what kinds of data that could be used in the validation process for performance assessments. Relevant empirical studies are cited in relation to most of the validity aspects, in order to further clarify either the meaning of the aspect, or how data can be used to support the validation process.

The *content aspect* determines content relevance and representativeness of the knowledge and skills revealed by the assessment. This is one of the traditional aspects of validity, which is often evaluated by means of experts' judgments (Miller & Linn, 2000). In any case, evidence for this aspect of validity should be grounded in a specification of the boundaries of the domain to be assessed, which in educational settings could be performed via task- and curriculum analyses (Messick, 1996).

Domain coverage is not only concerned with traditional content however, but also covers the "thinking processes", or the reasoning, used during the assessment. Such reasoning should, in an authentic assessment, ideally be the same as applied by professionals in the field when solving similar problems. The *substantive aspect* therefore has to include theoretical rationales for, and empirical evidence of, students actually using this reasoning when performing. Messick suggests that such evidence could consist of "think-aloud" protocols or correlation patterns among partial scores, and

Miller (1998) argues that expert judgments are needed, just as in the case of content validity.

Verbal protocols or small-scale interviews, together with observations of student performance, have been used to investigate the complexity of science assessments in a couple of studies. In a study by Hamilton, Nussbaum, and Snow (1997), it was shown that two very similar tasks, thought to assess the same reasoning, actually differed as to how the students solved them. This was due to students prior experiences with levers (such as the seesaw), as opposed to pendulums, which helped them find a correct explanation for the phenomenon in question. In another study, by Baxter and Glaser (1998), it was found that while some tasks required in-depth understanding of subject-matter knowledge, some tasks could be interpreted by the students at a surface level instead of at the intended level.

Interestingly, in the study by Baxter and Glaser (1998), the authors also identified tasks that elicited the appropriate reasoning, but where the scoring system was not aligned with task demands. In such cases, students did not get rewarded for their proper engagement in the task, or they could bypass the complexity of the task and still get high scores. This points to the fact that, not only do the tasks have to be consistent with the theory of the domain in question, but the scoring structure (such as the assessment criteria) must also follow rationally from domain theory. This is called the *structural aspect* of construct validity (Messick, 1996). Since the scoring method aids in defining the boundaries of the domain, criteria and standards should, according to Miller (1998), also be reviewed by experts in the field.

Messick singles out two aspects of validity in relation to the need for results to be generalizable to the domain in question, and not be limited to the particular sample of assessed tasks: the *generalizability* and the *external aspects*. The first refers to the extent to which score interpretations generalize across groups, occasions, tasks, etc., while the second concerns the relationship of the assessment score to other measures relevant to the domain being assessed. According to Messick, evidence of generalizability is the generalizability across occasions and assessors (i.e. the reliability concerns discussed previously, which will not be commented on fur-

ther here). Several researchers (e.g. Linn et al., 1991; Gielen, Dochy, & Dierick, 2003) suggest that the concept of reliability should be replaced by generalizability, and that generalizability theory (see Shavelson & Webb, 1991) should be used for investigating the degree to which performance assessment results can be generalized.

Suggested evidence for the external aspect of construct validity includes convergent correlation patterns with measures of the same domain, as well as discriminant evidence showing a distinctness from other domains. This is very problematic, however, since it presupposes that content and method could actually be separated, whereas current theories on learning assume that learning is *situated*. This means that learning, thinking, and acting are inseparable parts of an activity, and when students learn subject-matter facts or concepts, they are at the same time learning how to think and act in a certain community of practice (Lave & Wenger, 1991; Säljö, 2005; Wenger, 1998). As a consequence, content and method would be inseparable in such a perspective. This view is also supported by empirical studies comparing different assessment methods addressing the same content. Miller (1998) presents results from a study including multiple-choice items, both short and extended written responses to questions, as well as hands-on tasks. Correlations between the different methods showed that the method had a strong effect, since correlations were of moderate size only when the same method was used (i.e. between different multiple-choice items, between short- and long-answer questions, and between different performance tasks). When dissimilar methods were used, the correlations were close to zero, and these findings were consistent across several educational levels. Similar findings are reported by Dunbar et al. (1991), using data from different studies investigating writing performance, and by Ruiz-Primo, Li, Ayala, and Shavelson (2004) comparing different performance tasks in science. If correlations are used in order to gather evidence for the external aspect of construct validity, it would thus seem that the assessment methods compared should be of similar kinds. For performance assessments, this means that performance tasks should be compared to performance tasks only, and not, for instance, to multiple-choice or short-answer items.

The focus of the *consequential* aspect of construct validity is the intended and unintended consequences of assessments and the impact these assessments have on score interpretation[3]. An example of intended consequences could be changes in the instructional practice of teachers brought about by national assessments, while unintended consequences might include bias in the assessment (Messick, 1996; Miller, 1998). As an example of studies investigating intended consequences, Miller (1999) examined the perceptions of teachers about the consequences of state mandated performance assessments. In this study, most teachers agreed that the performance assessments had a positive influence regarding the alignment of instruction towards the curriculum.

Regarding bias towards population subgroups, Linn et al. write that:

> It would be a mistake to assume that shifting from fixedresponse [sic] standardized tests to performance-based assessments would obviate concerns about biases against racial/ethnic minorities or that such a shift would necessarily lead to equality of performance. Results from the National Assessment of Educational Progress (NAEP), for example, indicate that the difference in average achievement between Black and White students is of essentially the same size in writing (assessed by open-ended essays) as in reading (assessed primarily, albeit not exclusively, by multiple-choice questions). (Linn et al., 1991, p. 8)

The authors give other examples as well, where prevailing conditions of bias have not changed with the introduction of performance assessments, and they conclude that the question of fairness will probably be as pressing for performance assessments as for traditional tests.

In educational assessment, the most important consequence of assessment is student learning – the very raison d'être for all educational activities. It has been shown that assessment strongly affects student learning (e.g. Struyven et al., 2005), and this consequence

[3] By incorporating value implications and social consequences in the framework, some controversy has followed. This is because several authors oppose to including such aspects (even if acknowledging their importance) into the concept of validity (see e.g. Messick, 1998; Popham, 1997).

is sometimes unintended, as when assessment steers student learning towards surface approaches to learning. At other times, assessment is designed to actively affect student learning, as in formative assessment. When focusing on more complex forms of knowledge and performance assessments, it is thought that open-ended tasks are needed in order to elicit students' higher-order thinking. This is also because students are supposed to learn – through the backwash effect – complex ways of thinking and problem-solving skills, if these complex performances are indeed assessed. Whether, and how, this actually occurs, is an empirical question, which could be addressed under the heading of "consequential validity". This issue, however, will be the main focus of the next section (*The question of student learning*), and it is therefore not further elaborated on here.

The question of interest at this stage is how the problem of validity of performance assessments can be approached. Even though there seems to be no easy or single answer to this question, it is suggested that the validation process could be guided by a more comprehensive framework, instead of only focusing on isolated aspects of validity, such as content validity. By giving attention to the different aspects of validity, and by providing theoretical arguments and empirical data to support each of these aspects, such an approach could potentially aid in making an assessment more valid for its intended purpose. In educational settings, for instance, this could be achieved: by showing that the assessment is aligned with learning objectives; by showing that students use the kind of reasoning that was intended; and also that the assessment structure actually rewards students who engage in these processes, as opposed to those students who for some reason manage to bypass the complexity of the task. Other data could show that assessment scores generalize across tasks and assessors, and, if possible, that the students' scores correlate with other (similar) measures of the same domain. Investigations of "consequential validity" could search either for potential bias or for positive consequences, such as student learning, or both. Whether the assessment is to be considered valid or not, however, has to be decided by considering these data and arguments in relation to the intended purpose of the assessment. Also, this decision will have to be re-evaluated if

changes are made in the context, since such changes most probably affect the validity aspects.

The problem of credibility: Conclusions

Assessments have to be credible. Using performance assessments could in this respect be problematic, since research has shown that students' performance on complex and open-ended tasks varies considerably, as do sometimes the assessments by different assessors. Furthermore, assessment of complex tasks could easily be contaminated by domain-irrelevant factors, making it easier or more difficult for some students to receive high scores independently of their subject-specific knowledge or skills. In the light of these findings, it might be tempting to insist on the continued use of "traditional testing". That would be unfortunate, however, since written tests can only measure a limited part of students' knowledge and skills, and other modes of assessment are needed in order to "capture" the rest. This means that the use of written tests alone would clearly suffer from "domain under-representation"; something that would be especially true in a profession-oriented education. If examinations in such educational settings were based on only a limited part of the educational objectives, such as the aspects that are measurable by written tests, the grade/degree achieved would not represent the full scope of the competencies at stake. Therefore, instead of arguing for the continued use of traditional tests, it would make more sense to let the educational institutions using performance assessments for summative purposes demonstrate that these assessments are both reliable and valid. This process could in turn be facilitated by the use of a comprehensive framework of validity, such as Messick's, which also addresses the classic reliability issues.

In order to facilitate the design and inclusion of performance assessments, certain general principles can be extracted from the discussion above:

- to use detailed scoring protocols,
- to use sampled responses which exemplify the points on the scoring scale,
- to train the assessors,
- to use more than one task in the assessment,
- to closely specify the domain/objectives to be assessed,
- to make sure that the tasks elicit the performance to be assessed,
- to align the assessment criteria with the domain/objectives to be assessed, so that domain-irrelevant performance is not rewarded, and
- to check for possible bias.

The question of student learning

Research has shown that assessment can have a strong influence on student learning (Struyven et al., 2005). For instance, in an old but illustrative study by Säljö (1975), 40 students were divided into two groups. Both groups received the same task, to read a textbook and answer some questions after each chapter, but the groups were asked different kinds of questions. One set of questions concentrated on more or less verbatim reproduction of the text, the other set asked for deeper understanding. When they had finished the whole book, both groups had to answer both kinds of questions, and they also had to summarize the book in a few sentences. Afterwards, the majority of the students stated that they had directed their learning strategies towards what they believed was required of them, which in this case was affected by the kind of questions they had received after each chapter. This effect was also clearly visible in the students' results, where those students who had received questions asking them to reproduce the text had adopted a surface approach to learning.

From these and similar results, it might be assumed that the "backwash effect" of assessment could be used in a positive way. If the assessments were changed towards assessing understanding or complex performance (i.e. using performance assessment), students would then supposedly adapt their learning strategies towards

deep-learning approaches. Unfortunately, the matter does not seem to be quite so simple. In Säljö's study, in the group who received questions asking for deeper understanding of the text, not all students adopted a deep-learning approach. Some did, while others did not, and the latter instead managed to solve the tasks in an instrumental way. It would thus seem that it is quite easy to get students to adopt a surface approach to learning, but much harder to help them acquire a deep-learning approach (Gijbels & Dochy, 2006; Marton & Säljö, 1984; Struyven et al., 2005; Säljö, 1975; Wiiand, 1998).

According to Marton and Säljö (1984), motivational factors can offer an explanation as to why some students did, or did not, use a deep-learning approach. Intrinsic motivation (i.e. that the students want to understand) can be assumed to be linked to deep learning, and the fact that some students adopted a deep-learning approach could therefore be linked to their interest and desire to understand the particular text. This, however, would imply that students' learning approaches are extremely sensitive to topic. Also, it is hard to know if you are interested in a text before you actually read it, and then the learning approach would perhaps have to be changed under way.

To investigate whether there were differences in experiences of learning which could affect preferences for learning approaches, a follow-up study was conducted, where interviewees were asked about their perceptions of learning. Results showed that there were different perceptions of learning among the participants, and that these perceptions could be linked to surface- and deep-learning approaches. It was thus suggested that those who had a more developed conception of learning could become aware of their own learning, and in this way be able to adapt their learning approaches to different tasks (Entwistle & Peterson, 2004; Marton & Säljö, 1984).

However, subsequent research by Entwistle and Ramsden (1983) on students' everyday studying, indicated that students' approaches to learning tended to be affected by assessment demands, rather than representing characteristics of the individual learners. Consequently, an additional category besides surface- and deep-learning approaches was introduced, called the "strategic orientation" (cf.

"achieving approach" in Biggs, 1987)[4]. The aim of this approach is to get high marks, and thus learning is only viewed as the means of the educational enterprise, not the end[5]. Taking this view, another reasonable explanation as to why some students did, or did not, use a deep-learning approach, could be differences in their perception and interpretation of the assessment context and of what was required of them.

That the perceived context, and not necessarily what the context is "really like" in an objective sense, is important for students' learning strategies, has been shown by Fransson (1977). In his study, some students, due to their perceptions of the context, adapted their approaches to learning towards an expected mode of assessment, although this assessment had not been announced. Also, in a study by Segers, Nijhuis, and Gijselaers (2006), students in a problem-based course adopted less deep-learning and more surface-learning approaches than students in a more conventional course, which was quite contrary to expectation. Since students in the problem-based course did not differ in their perceptions of the assessment demands, as compared to the students in the more conventional course, there may have been other contextual factors (such as workload) that induced the students to adopt a surface approach. Therefore it is important to make a distinction between the context as seen from the outside, for instance as defined by teachers or researchers, and how it is perceived by the students (Entwistle & Peterson, 2004). Furthermore, this means that the students need to be *aware* of what is expected of them, if they are to adapt their learning approaches (or perhaps more appropriately learning "strategies"), to the assessment requirements. Otherwise they will be guided by things like personal motivation and/or prior experiences of assessment (Entwistle, 1991; Marton & Säljö, 1984; Segers et al., 2006; Struyven et al., 2005).

Student awareness of the purpose of the assessment and the assessment criteria is often referred to as *transparency.* Transparency has been shown to be important for the students, for instance since

[4] There are also a number of other related concepts in the literature, such as "study orchestration". See Lonka, Olkinuora, and Mäkinen (2004) for a more thorough discussion on this topic.

[5] This orientation (or approach) has been confirmed in several studies, and it has also been shown that students who prefer deep-learning approaches indeed tend to receive lower grades than students who adopt the "strategic orientation" (Svingby, 1998).

many university students believe that not knowing what is expected of them has a very negative impact on their learning (Wiiand, 2005).

Frederiksen and Collins (1989) argue that – if tests are to be considered valid – students should be able to assess themselves with nearly the same accuracy as the test developers. In their appeal for "systematically valid testing", they write that the validity of tests should include the effects these have on instruction and learning (cf. Messick's consequential validity). If a test is to be "systematically valid", it should lead to changes in instruction that promote learning of the same skills that the test was intended to assess, which is another way of arguing for the positive "backwash effect" described previously. According to Frederiksen and Collins, however, it is not satisfactory to design tests or assessments which are *supposed* to lead to positive changes. Instead, all assessments should comprise explicit means to enhance these effects. The authors suggest that the following features should be included: (1) practice in self-assessment, (2) repeated testing, (3) feedback on test performance, and (4) multiple levels of success. These features have considerable overlap with points made by other authors (e.g. Black & Wiliam, 1998b; Wiggins, 1998), and are also grounded in empirical findings showing that practice in self-assessment and feedback, for instance, can be quite effective means to raise educational standards (Black & Wiliam, 1998a).

Not everybody agrees that transparency might be an effective way to help students improve their work. For instance, Dysthe, Engelsen, Madsen, and Wittek (2008) argue that criteria development should be an ongoing process of negotiation, rather than having students accept the "authoritative word" (i.e. explicit criteria) of the teacher. Furthermore, Messick (1996) writes that transparency may be counterproductive, since it could impede originality and innovation. These objections to transparency, however, conceal some underlying assumptions, which are not necessarily true:

1. The assumption that there is no established consensus about quality criteria in professional communities. On the contrary, there is research indicating that such commonly accepted criteria exist, even in domains regarded as mainly tacit, although the criteria may not be articulated (Lindström, 2001, 2008). Involving the student in the negotiation of criteria might, however, be appropriate in novel domains or in situations where there is no existing consensus (such as in portfolio assessment).
2. The assumption that some performances, such as creative performance and innovative work, are not possible to assess. On the contrary, as Lindström (2006) convincingly has shown, it is quite possible to assess for instance creativity.
3. The assumption that individual creativity is prior to societal conventions. On the contrary, it could be argued that artists and authors are sometimes given too much credit for the novelty in their work, since they actually work within a network of social conventions. High quality work is not created *ex nihilo*, but from influences by predecessors, and what is considered "good" is also decided by social conventions (Wertsch, 1998).

In summary, students need to be aware of what is expected of them if they are to adapt their learning strategies to the assessment requirements. Provided that transparency is accepted as a means for enhancing this positive backwash effect of assessment, practice in self-assessment, repeated testing, feedback on test performance, and multiple levels of success could be included to facilitate student learning. However, the points made by Frederiksen and Collins (1989) would still have to be further clarified, before any general principles could be extracted to guide the process of designing valid performance assessments for formative use; for example: "What kind of feedback should be used?", "How should the self-assessment be carried out?", and "How can multiple levels of success be used to let the students strive for higher standards?". An attempt to answer these questions is made in the following sections.

Feedback

Feedback can take on many shapes and be used for a variety of reasons. For instance, in a study investigating teacher feedback, Tunstall and Gipps (1996) discriminate between evaluative and descriptive feedback, which are seen as endpoints in a continuum. At the evaluative end, feedback is either positive or negative, and the feedback is given according to explicit or implicit norms for socialization purposes. At the other end of the continuum, descriptive feedback is focused on achievement or improvement related to actual performance. In order for the feedback to support learning, this distinction between evaluative and descriptive feedback is clearly important, since evaluative feedback either has no effect on learning, or even negative effects, while descriptive feedback tends to give positive effects (Black & Wiliam, 1998a).

This tension between feedback directed towards performance or self, is further corroborated in a large meta-analysis on feedback interventions by Kluger and DeNisi (1996), including more than 600 effect sizes. According to these authors, feedback interventions can yield large effects if the feedback is directed towards performance instead of the self. But even if there are substantial positive effects during the intervention, these effects disappear, or become negative, if the feedback is removed.

For the feedback to have continuous positive effects, according to Sadler (1989, 1998), it is necessary for students *themselves* to be able to check the quality of their own work against assessment criteria or standards. This, however, requires that the students understand what high-quality work looks like. Furthermore, they must have the necessary skills to compare their own performance with work of higher standards, and adjust their performance in order to reduce the gap between own performance and the performance aimed for (see also Wiggins, 1998). This points to the strong link between feedback and self-assessment, as it is not specified who is giving the feedback – it could be the teacher, a peer, an artifact or the student herself – and there is a clear reference to student agency.

In relation to this discussion on student agency and self-assessment, Tunstall and Gipps (1996) also distinguish between different modes of descriptive feedback. In ”specifying” feedback

the teacher tells the student what needs to be done in order to "close the gap". Viewed from Sadler's perspective, this kind of feedback would not promote learning, but instead make the student dependent upon the teacher's expertise. "Constructing" feedback, on the other hand, differs from "specifying" feedback by sharing responsibility for the assessment between teacher and student, both parties discussing and agreeing on what needs to be improved. In such a model, the teachers share not only their professional judgments, but also their interpretations of quality and standards, and the students become active participants in the assessment process, which could potentially give them the skills Sadler and Wiggins are arguing for (Gipps, 2001; Sadler, 1989; Tunstall & Gipps, op. cit.; Wiggins, 1998).

In conclusion, feedback should be descriptive and task related in order to support learning. Furthermore, feedback should not only be handed over to the student, telling her what needs to be done. Instead, the student must learn to construct her own feedback, which could be done in interactivity with the teacher and/or by practice in self-assessment.

Self-assessment

There exists considerable empirical research on self-assessment, and this research is quite evidently influenced by the scientific paradigm within which it is conducted. Consequently, many studies working within the psychometric tradition focus on the quantitative agreements of grades by students and their teacher (see Falchikov & Boud, 1989; Boud & Falchikov, 1989). When analyzing these studies at a meta level, some interesting findings appear. One such finding is that even though senior students are sometimes quite skillful at assessing themselves in the subject they have been studying for some time, they are not more skilled in self-assessment than novices when they self-assess in subjects new to them (Boud & Falchikov, op. cit.). This indicates that self-assessment is not a generic ability we are born with, but rather a contextualized skill that can be learned and improved by practice and feedback, a conclusion also supported by other reviews of empirical research on self-assessment (Dochy, Segers, & Sluijsmans, 1999; Topping, 2003). As a consequence, it would make sense to let the students

practice self-assessment embedded in subject specific (or professional) activities in an authentic manner.

Regarding the issue of student learning, Dochy et al. (1999) have published a review on self-assessment in higher education which is not limited to quantitative comparisons between students and teachers, but also includes application of self-assessment in natural settings and different kinds of instruments used to investigate students' self-assessment. From the studies reviewed, they conclude that self-assessment, and different combinations of peer-, self- and co-assessment with the teacher, can have several positive effects beyond improved agreement with teachers' assessment. Examples are increased confidence, increased awareness of the quality of their own work, increased reflection on own performance, responsibility for learning, and increased satisfaction. Instruments used to estimate self-assessment skills were various Likert scales, but also interactive systems, letters, portfolios, and audio recordings.

Another recent review on self-assessment was performed by Topping (2003), focusing partly on questions of reliability and validity, but also on effects of self-assessment in schools and higher education. The empirical support for the development of meta-cognitive skills[6] due to training in self-assessment is, according to Topping, small but encouraging. The results indicate that self-assessment can promote: (1) strategies for coping with own learning, (2) self efficacy, (3) deep learning, and (4) gains on traditional summative tests. He also notes that the effects are at least as good as those from more conventional modes of assessment, and often better, but that these effects might not appear immediately.

In conclusion, since research does not support the notion of self-assessment skills being generic and context independent, students should practice self-assessment embedded in subject specific (or professional) activities in an authentic manner. Furthermore, there are several studies indicating increased meta-cognitive awareness as a result from practice in self-assessment, and these results are not confined to certain contexts or instruments, but come from a wide range of educational settings.

[6] For a discussion on different types of meta-cognitive knowledge and skills, see for example Flavell, Miller, and Miller (1993).

Multiple levels of success

The idea that criteria can be formulated in order to assess the qualities valued within a certain community of practice has been proposed by Dewey (1934/1980) among others. According to Eisner (1991), the use of criteria can facilitate the search for qualities that can not be measured quantitatively, and the criteria also make it possible to estimate these qualities. This as opposed to "standards", which is a term with many different meanings, but in Dewey's terminology means a quantitative measurement. But the question of "how much?" is, again according to Eisner, not very interesting from a learning point of view. The interesting question is "how good?". To assess by the use of criteria means that you have to be able to motivate and to argue for the assessment, this in turn requires an understanding of the criteria in relation to the particular community of practice. Thus it is much more complex to assess by the use of criteria than by the use of (quantitative) standards. Still, criteria have the potential of improving instruction and learning, which (quantitative) standards do not have.

Consequently, when providing "multiple levels of success", as suggested by Frederiksen and Collins (1989), the use of quantitative standards would not seem appropriate. Instead, "standards" as referred to by Sadler (1987, 1989) should be examples of performance which differs in *quality*. Wiggins (1998) states that: "A true standard /…/ points to and describes a specific and desirable level or degree of exemplary performance – a worthwhile target irrespective of whether most people can or cannot meet it at the moment" (pp. 104-105).

In conclusion, in order to support student learning, standards should not be points along a numerical scale, but examples of performances which differ in quality.

The question of student learning: Conclusions

The purpose of using performance assessments to facilitate student learning is twofold. First, open-ended tasks are thought to be needed in order to elicit students' higher-order thinking. Secondly, since assessment has been shown to direct students' learning, the backwash effect could potentially be used in order for students to learn complex ways of thinking and to acquire problem-solving

skills. As has also been shown, however, changes in the assessment do not automatically make students adopt appropriate learning strategies. What seems to be of crucial importance is how the students perceive and interpret the assessment context and requirements. It is thus argued that, by making the purpose of the assessment and the criteria explicit, the students become aware of the expectations, and can then consciously adapt to the assessment demands. The students have to be able to see that, for instance, using deep approaches to learning is required of them, if they are to succeed on the assessment.

In the discussion above, there is still the embedded assumption that students somehow know which learning strategies are appropriate. When using well-known modes of assessment (such as multiple-choice questions or essays) in a way recognized by the students, they seem to adapt their learning strategies quite successfully to the assessment demands, as shown for instance by Scouller (1998). But when using the same methods in a different way, or when confronted with novel forms of assessment, it is doubtful whether students will know how to adapt their learning strategies (Segers et al., 2006). This is seen, for instance, in a study by Gijbels, Van de Watering, Dochy, and Van den Bossche (2005), where multiple-choice questions were used to assess problem-solving instead of recall. The results from this study show that many students received low scores for both deep- and surface-learning strategies, and there were no relationships between employment of deep-learning strategies and higher assessment outcomes. Furthermore, as shown in a study by Lindblom-Ylänne (2003), several law students realized that their approaches were not suitable for the subject they were studying, but they did not know how to change them in an appropriate direction. In order to avoid such a mismatch, or "dissonance", of student learning and assessment demands, it is suggested that all assessment have in-built features for the fostering of appropriate learning strategies, such as practice in self-assessment, repeated testing, feedback on test performance, and multiple levels of success. These features could be used as general principles to guide the process of designing performance assessments for formative use, if:

- the self-assessment is embedded in professional (or subject specific) activities in an authentic manner – as opposed to being decontextualized and generic,
- the feedback is descriptive (task related) and "constructed" – as opposed to evaluative (self related) and "specifying", and
- standards are examples of performance which differ in quality – as opposed to points along a numerical scale.

STUDY I: THE USE OF RUBRICS

The previous chapter took as a starting point two of the problems with introducing performance assessment in higher education in general. These problems were: (1) whether assessment of complex performance can be carried out in a credible and trustworthy manner, and (2) whether the assessment actually affects student learning. At the end of the chapter, a set of assumptions about how to deal with these issues were reached. Most of these assumptions are based on relevant empirical research. The assumptions could possibly be used as general principles, or guidelines, for designing performance assessments in teacher education, but in their current form they are still quite non-specific.

In order to investigate the problems of performance assessments further, a tool for adding to the quality of performance assessments is introduced in this chapter: the scoring rubric. Although the term "rubric", is used in several different ways in the literature, a widespread definition is that the rubric is a scoring tool for qualitative rating of authentic or complex student work. A characteristic of rubrics is that they include criteria (i.e. important dimensions of performance) as well as standards for those criteria (i.e. qualitatively distinct levels of performance). By making the criteria and standards explicit, the rubric tells both instructor and student what is considered important and what to look for when assessing (Arter & McTighe, 2001; Busching, 1998; Perlman, 2003). This holds for both high-stake summative assessments as for formative assessments. A simplified example of a scoring rubric is seen in Figure 1.

Assessment of…	*Lowest level*	*Highest level*
Formulation of a hypothesis Description: A hypothesis is…	The hypothesis is not testable.	The hypothesis is testable.

Figure 1. *A simplified example of a scoring rubric. If the educational objective is for the students to have "knowledge about the scientific method", then one aspect of this is to be able to formulate a hypothesis (a performance). However, all hypotheses are not of equal quality. For example, a hypothesis has to be testable in order to be of scientific value. By expressing different quality levels in the rubric, students' performances can be assessed, but the standards also express what is considered high-quality performance. The next step in the construction of this particular rubric could be to add either more criteria for the formulation of a hypothesis (e.g. whether the hypothesis is relevant and backed up by previous research, and/or that it is expressed as simply as possible), or adding other aspects to be assessed (viz "designing an experiment", "collecting data", etc.).*

Several benefits of using rubrics are mentioned in the literature. One widely cited effect of rubric use is the increased consistency of judgment when assessing performance and authentic tasks. Rubrics are assumed to enhance the consistency of scoring across students, assignments, as well as between different assessors. Another frequently mentioned positive effect is the possibility to provide valid judgment of performance assessment that cannot be achieved by means of conventional written tests. It seems that rubrics offer a way to provide the desired validity in assessing complex performance, without sacrificing the need for reliability. Another important effect of rubric use proposed, is the promotion of learning. It is assumed that the explicitness of criteria and standards are fundamental in providing the students with quality feedback, and ru-

brics can in this way promote student learning (Arter & McTighe, 2001; Wiggins, 1998).

When comparing these statements about rubrics with the problems of introducing performance assessments in higher education explored in the previous chapter, it would seem that the use of rubrics could actually remedy these problems to some extent. However, even though the benefits mentioned may seem plausible, research evidence to back them up is needed. Therefore a review of available research literature on rubrics relating to these statements has been performed, where it was investigated whether evidence could be found on the effects of rubrics in high-stake summative, as well as in formative, assessments (Jonsson & Svingby, 2007).

Research questions

The review aimed to answer the following questions:

1. Does the use of rubrics enhance the reliability of scoring?
2. Can rubrics facilitate valid judgment of performance assessments?
3. Does the use of rubrics promote learning and/or improve instruction?

In the following, results from the review are presented in a condensed form and discussed in relation to the problems of credibility and student learning. Further details on results, procedure, and data are available in the article[7].

[7] It should be noted that the review is not a meta-analysis of the empirical data, due mainly to the heterogeneity of the material, where some rubric conditions were only represented by a few cases.

Does the use of rubrics enhance the reliability of scoring?

In the previous chapter it was shown that there are three main factors affecting the reliability of performance assessments (i.e. assessors, tasks, and occasions). When investigating the scoring reliability of rubrics, assessors and occasions are both of importance. These factors are included in the concepts of inter-rater and intra-rater reliability (i.e. the variation between different assessors and the difference between different occasions for the same assessor), respectively.

Studies focusing on intra-rater reliability were relatively few, but they indicated that rubrics can aid assessors to achieve high internal consistency when scoring performance tasks. These results thus support Brennan's (2000) remark that the occasion aspect of reliability might make a relatively small contribution as compared to other sources of error.

There were considerably more studies reporting on inter-rater reliability as compared to intra-rater reliability. Furthermore, the majority of the results reported on assessor reliability did not reach the criteria set for reliable scoring, such as 70 percent agreement for exact agreement or .70 for correlation of scores among assessors.

Several findings in the review support the fact discussed in the previous chapter: when all students do the same task, the reliability will most likely be high. But when students perform different tasks, choose their own topics or produce unique items, then reliability is often lower (Brennan, 2000). An example is that extraordinary high correlations of assessor scores are reported by Ramos, Schafer, and Tracz (2003) for some items on the "Fresno test of competence" in evidence-based medicine, while the lowest coefficients are for essay writing. Tasks like oral presentations also produce relatively low values, whereas assessment of for example motor performance in physical education (Williams & Rink, 2003) and scenarios in engineering education (McMartin, McKenna, & Youssefi, 2000) report somewhat higher reliability.

The question guiding this dissertation, however, is the inclusion of performance assessments in higher education, where the reliability needs to be quite high, even if students perform different tasks,

choose their own topics or produce unique items. As there are several factors influencing inter-rater reliability reported in the literature, this information can be used to get a picture of how to make rubrics for performance assessments more reliable:

1. Agreement among assessors seems to be increased by the use of benchmarks[8] and by training.
2. Analytical scoring, as compared to holistic[9], is often more reliable.
3. Topic- or task-specific rubrics are likely to produce more generalizable and dependable scores than generic rubrics.
4. Augmentation of the rating scale (for example that the assessors can expand the number of levels using + or - signs) seems to improve certain aspects of inter-rater reliability, although not consensus agreements. For high levels of consensus agreement, a two-level scale (for example competent vs. not competent performance) can be reliably scored with minimal training, whereas a four-level scale is more difficult to use.
5. Two assessors are often enough to produce acceptable levels of inter-rater agreement.

In the previous discussion on the problem of credibility, it was concluded that detailed scoring protocols should be used in order to increase the reliability of scoring, and rubrics are a kind of scoring protocols that could be used in this respect. Even though many rubrics do not provide sufficient reliability if used on their own, findings indicate that the reliability can be increased by means of the strategies mentioned above (point 1-5). For instance, as proposed in the discussion on credibility, sampled responses which exemplify the points on the scoring scale could be used. In a study by Denner, Salzman, and Harris (2002), assessment with and without such benchmarks were compared. While assessing without bench-

[8] The distinction between ”benchmarks” and ”exemplars” is not always made clear in the literature. However, ”benchmarks” are most often used to denote *written descriptions* of the levels in a rubric, while ”exemplars” are mainly used to denote *examples of performance* (authentic or fictional) representing the different levels in the rubric (cf. Sadler, 1987). ”Model answers”, on the other hand, is a term used to denote *ideal responses* (see e.g. Huxham, 2007).

[9] In holistic scoring, the assessor makes an overall judgment about the quality of performance, while in analytic scoring, the assessor assigns a score to each of the dimensions being assessed in the task.

marks did not reach acceptable levels of reliability, scoring with benchmarks did. A similar trend is visible for analytical versus holistical scoring and topic-specific versus less specified conditions. Augmented rating scales, in combination with analytical scoring, could also be mentioned, but increase in reliability by augmentation comes with a prize, as the consensus agreement seems to drop considerably when extending the rating scale.

In conclusion, rubrics could be used to increase reliability, but they should be complemented with benchmarks and assessor training. Also, the rubrics should be analytic and topic-specific. Depending on the levels of reliability reached by using these strategies, two independent assessors could be used, but adding more assessors does not necessarily increase the reliability to any significant extent. Whether augmentation of the rating scale could be used or not, would depend on how the score is used, for example whether sub-scores are summarized to an overall score or whether decisions of "competent" versus "not competent" performance are made at the criterion level.

Can rubrics facilitate valid judgment of performance assessments?

Most studies reviewed claim to have support for the validity of the rubric used. Several studies report that the rubric used has been validated for content validity by experts, and the scores produced have been checked for correlation with other measures, both internal and external. Researchers have performed factor analyses to reveal the underlying theoretical constructs, or investigated the alignment of guidelines, standards, and rubrics. However, only one study in the review (Gearhart, Herman, Novak, & Wolf, 1995), has used a more comprehensive framework for the validation process. It is therefore relevant to ask what it means when a rubric has been shown to have (for instance) content validity, and no other aspect of validity has been addressed. It could mean that content knowledge is properly assessed, while other dimensions, like reasoning or generalizability, are not. It could also mean that there is no alignment between objectives and assessment, or that there are

severe social consequences or bias. All these factors threaten validity and might produce unfair results.

On the issue of reliability, it could be assumed that, since a rubric is a regulatory device, scoring with a rubric is probably more reliable than scoring without one. It cannot, however, be concluded that scoring with a rubric is probably more *valid*. Just by providing a rubric there is no evidence for content-representativeness, fidelity of scoring structure to the assessed domain, or generalizability. Nor does it give any convergent or discriminant evidence to other measures.

But when considering the validity of rubrics in relation to the general principles arrived at in the discussion on credibility, the use of rubrics could assume the role of specifying the domain to be assessed. If the criteria are developed in accordance with educational objectives, the rubric defines what is to be assessed, thus potentially minimizing the influence of domain-irrelevant factors. Also, by only rewarding performance sought for, the use of a rubric could help eliciting this performance.

In conclusion, even though the use of a rubric does not automatically provide validity to the assessment, rubrics could play the role of specifying the domain to be assessed, and in this way aid in minimizing domain-irrelevant variance. Also, by only rewarding domain-relevant performance, the use of a rubric could help in eliciting the performance required. To be able to support valid assessment in this way, the rubric would have to be developed in accordance with educational objectives or domain theory, and be validated with the aid of a comprehensive framework of validity.

Does the use of rubrics promote learning and/or improve instruction?

Advocates for the use of rubrics for formative assessment assume that rubrics can promote student learning, as well as lead to positive changes in instruction. This could be done in several different ways, for example in either a teacher- or a student-centered way. Regarding the former, the rubric could be used by the teacher to enhance the alignment of learning, instruction, and assessment,

something that is often referred to as "constructive alignment" (Biggs, 1996). In a more student-centered approach, the rubric could be shared with the students (see the discussion on "constructing feedback" in the previous chapter), aiming for self-regulation (Wiggins, 1998). Both of these strategies are represented in the articles reviewed, and the majority report their findings as student improvement and/or perceptions of using rubrics.

Perceptions of using rubrics

A major theme in the comments from both teachers and students is the perception of clarified expectations, or transparency. Rubrics indicate what is important and thereby provide clarity and explicitness to the assessment, and this is deemed positive by students and teachers alike. Besides transparency, other benefits of rubrics perceived by the teachers are the encouragement of reflective practice and that rubrics can give teachers more insights into the effectiveness of their instructional practices. Also, the concrete nature of rubric criteria provides information for feedback as well as making self-assessment easier.

It would thus seem that the use of rubrics has the potential of promoting learning and/or improving instruction, at least as perceived by the teachers and students using them. The way in which rubrics support learning and instruction is by making expectations and criteria explicit, which also facilitates feedback and self-assessment. As with valid assessment, however, student learning should probably not be expected to improve just by introducing a rubric. The assessment also needs to have in-built features for the fostering of appropriate learning strategies, such as practice in self-assessment, repeated testing, feedback on performance, and multiple levels of success (Frederiksen & Collins, 1989). A rubric could, if designed according to the principles for formative assessment arrived at in the previous chapter (such as being task related and having quality standards), facilitate both feedback and self-assessment, which is also noted in some of the studies on rubrics.

Interpretation of criteria

To interpret criteria is difficult, especially for novices (like students), which is illustrated by Orsmond and Merry (1996). In one

of their studies, criteria like "a clear and justified conclusion" were so alien to the students that they were not able to recognize such a conclusion, even though it was described to them by the instructor. In another study by the same authors, students worked in pairs on a performance task. It was assumed that the students would discuss and come to a shared understanding of the criteria, but this assumption could not be confirmed (Orsmond, Merry, & Reiling, 1997). Instead, this and other studies point to the fact that students need practice, and/or sampled responses which exemplify the points on the scoring scale, in order to interpret the criteria and standards in the rubric.

The fact that practice is needed in order to interpret the criteria and standards operationalized in a rubric is demonstrated by studies comparing "rubric-only" conditions with the combination of a rubric and other instructional interventions. In a study by Duke (2003), students getting both a rubric and "cognitive-strategy" instruction, did better on essay writing than students who received only the rubric and brief explanations of the criteria. The author concludes that the combination of rubric and instruction proved to encourage students' self-regulatory behaviors. This study concords with a study by Toth, Suthers, and Lesgold (2002), in indicating that the combination of a rubric with practice in understanding the criteria might be needed in order for students to improve. These results are further supported by studies investigating student improvement in relation to rubric use.

Student improvement

The analysis of the studies reviewed shows that there can be quite dramatic effects on students' performance when a rubric is used as an assessment tool for learning. However, in line with the reasoning above, the rubric in most studies is combined with the possibility of learning how to use the rubric, for example through self-assessment.

An example where a marked positive effect on student performance is demonstrated, is in a study by Andrade (1999b). Students in a science class self-assessed their work with the assistance of a rubric. Results show that the treatment group considerably outper-

formed the control group (effect size = .99)[10]. Similarly, the students in a study by Brown et al. (2004) showed quite large improvements (effect size = 1.6). Here the rubric was used in a training program for writing, involving guidance in "meta-cognitive monitoring". In yet another study reporting on student improvement, the context was student laboratory write-ups. The writing was supported by a rubric as well as peer-editing sessions and self-assessment. On an average, the scores of the write-ups improved by 17 percent in this study (Mullen, 2003). The last example is a study by Schamber and Mahoney (2006), where the combination of writing assignments and the use of a rubric improved the scores in an assessment of critical-thinking skills by 41 percent. Although few, these studies indicate that rubrics might be valuable in supporting student learning. Some studies also show that students actually internalize the criteria in the rubric, making them their own, and use them while self-assessing (Andrade, 1999b; Piscitello, 2001).

There exist, however, also studies presenting results which are not as straightforward. Andrade (1999a) has conducted a couple of studies on writing performance, where the students were supported by a rubric and self-assessment, and she reports positive effects only for some of the groups investigated – and the results sometimes differ between the sexes.

To conclude, rubrics seem to be able to aid in improving student performance, but even though this summary of findings on the educational consequences of rubrics is promising, there are still some important issues to be considered.

First, providing students with a rubric does not automatically lead to student improvement. What also seems to be needed is some kind of practice, where the students become accustomed to the criteria. Examples which have shown to be successful (although not in all contexts) are "cognitive-strategy instruction", self- and peer assessment, and "meta-cognitive monitoring". It could be noted that all of these involve some kind of meta-cognitive activity, and aim for self-regulation.

[10] This could be compared to effect sizes for formative assessment in general, which are typically between .4 to .7 (Black & Wiliam, 1998a).

That some kind of additional intervention is needed also indicates the importance of "constructive alignment" (Biggs, 1996). As Segers et al. (2006) note, many assessment studies focus on changes in the assessment, while more or less neglecting the learning environment, which might possibly explain the low impact on students' learning. These authors also suggest that it is not the constructive alignment as designed by the teacher or the researcher which affects student learning, but the constructive alignment as perceived by the students. This means that assessment and instruction should focus on the same knowledge. The students should ideally be able to perceive this alignment and adapt their learning strategies accordingly. Working with rubrics, this does not necessarily imply that this alignment has to be expressed according to a specific learning theory (like constructivism), but only that the same criteria are used in both instruction and assessment, and that students are given the opportunity to understand the criteria.

Second, there are some ambiguities in the results. Not all studies demonstrated improved student performance, even if the rubric was complemented with self-assessment. There can be many potential explanations for these differences, but one important factor might be the time devoted to learning how to use and understand the rubric. In the study by Andrade (1999a), less than 40 minutes was spent on introducing and reviewing the rubric, which could be compared to the average of 27.25 hours spent teaching the students in the study by Brown et al. (2004).

Third, there are very few studies, representing a rather narrow field of performances and educational settings. For example, in several of the studies the performance investigated was some kind of writing, and most studies were performed in middle- or high schools. There was only one study at college level, and there were no studies in which students' actual behavior were assessed. This severely restricts the conclusions that may be drawn, and it could be questioned to what extent they generalize to higher education, in general, or to profession-directed education (where students need to learn how to act competently) in particular.

These issues will be further discussed in the next chapter, which explores how the empirical findings about rubrics can be used to design an assessment methodology intending to support student

learning as well as providing reliable and valid scoring ("educative assessment").

AUTHENTIC ASSESSMENT: PUTTING IT INTO PRACTICE

As was shown in the previous chapter, the introduction of rubrics can help remedy some of the problems which arise when introducing performance assessment in higher education. Conclusions from the review of empirical research on rubrics are outlined below.

A. In relation to student learning:

1. Rubrics can support learning and instruction by making expectations and criteria explicit, which also facilitates feedback and self-assessment.
2. Rubrics can aid in improving student performance.

In order to improve student learning, the rubric should be task related and indicate quality standards. Furthermore, self-assessment (or other meta-cognitive activity aiming towards self-regulation, by which the students become accustomed to, and understand, the criteria) is also needed.

B. In relation to credible assessment:

1. Rubrics can be used to increase reliability.
2. Rubrics can specify the domain to be assessed, and in this way aid in minimizing domain-irrelevant variance.
3. Rubrics can help in eliciting the performance required.

In order to have these effects, the rubric used should:

- be complemented with benchmarks and assessor training,
- be analytic and topic-specific,
- only reward domain-relevant performance,
- be developed in accordance with educational objectives or domain theory, and
- be validated with the aid of a comprehensive framework of validity.

Depending on the levels of *assessor reliability* reached by using the above mentioned strategies, two independent assessors could be used. Furthermore, the rating scale could be augmented, but whether this is an appropriate strategy would depend on how decisions are made on the basis of the score (i.e. whether decisions are based on an overall score or on individual criteria). An augmentation of the rating scale typically improves assessor reliability if estimated by correlation coefficients, but lowers reliability if estimated by consensus agreements. To increase *task reliability*, more than one task should be used in the assessment.

The above mentioned conclusions on how to remedy problems of credibility and to enhance student learning are supported by empirical findings. However, this research covers a wide range of contexts and research designs, from classroom studies to experiments in laboratory settings. Furthermore, there is great variation in research focus, in research questions posed, in analyses made, etc.

A problem with this great variation is that it makes it difficult to estimate the generalizability of the findings. For example, laboratory findings do not always generalize to classroom contexts (see for example Lundeberg & Fox, 1991), research performed in schools does not necessarily generalize to higher education, writing performance does necessarily generalize to other performances, etc. This means that the conclusions should be treated as a set of empirically based assumptions rather than evidenced theory, until more systematic research has been performed. Such research, linking to the reviewed studies and testing the assumptions in different

contexts, could provide more information as to the limits of these conclusions.

In this dissertation, the assumptions will be used in relation to instructional practice (i.e. not to experimental settings) in teacher education. This has been done by incorporating the assumptions into a performance-assessment methodology called the "Interactive examination", which is then used as a starting point for case-study research in a particular context.

The aim of this chapter is to describe the design of the assessment methodology used as well as the research design. This is done by:

- presenting the context in which the assessment methodology was implemented,
- discussing some of the specific problems associated with formulating criteria for (and assessing) teacher competency,
- presenting the assessment methodology (i.e. the "Interactive examination") as it was originally designed,
- presenting the assessment methodology as it has been used in this dissertation, highlighting the differences between the two,
- presenting the changes made between the pilot version of the assessment methodology and subsequent versions, and
- presenting the research methodology (sample, data, analyses, and limitations).

Context

The course in which the "Interactive examination" was implemented was called "Becoming a teacher", and this course was, at the time of implementation, held at the end of the first semester of the teacher-education program. As a whole, the course focused on learning by self-assessment and dialogues in net-based groups. This means that the students had the opportunity to practice self-assessment before the "Interactive examination", and also to reflect on own learning preferences, self efficacy, etc. (Folkesson & Svingby, unpublished manuscript). For example, the students assessed the social dynamics of their own net-based group work

(through Social Network Analysis; see Malmberg, 2006), as well as the quality of their own contributions to the dialogues (Malmberg & Svingby, 2004).

Since all students registered at the School of Teacher Education had to take the same courses during their first semester, regardless of their subject major or which school levels they were specializing towards, all students at the department in question ("Science-Environment-Society") had to take the "Interactive examination". This means that the student population doing the examination represented a wide array of personal backgrounds (for example native students who had gone through their entire "educational career" in Sweden, as well as those who had been born abroad and also had passed most of their schooling in another country; students from families where neither parent had any secondary education, and students from families where both parents held university diplomas). Also, the prospective professional settings differed (for example those students who wanted to work in preschool, or as leisure-time pedagogues, as well as those aiming for the upper-secondary level).

Objectives of the course assessed by the examination were that the student teachers should be able to *document, describe, and reflect on* [school] *students' situations*, as well as be able to *discuss the influence of different social and cultural conditions*. In addition, the students were expected, according to national goals for the entire Swedish teacher education, *to systematically use relevant scientific material as well as their own (and others') experiences as a basis for professional development* (see SFS 2001:23). Assessing these objectives during the first semester, and with this very heterogenic student population, meant that the examination tasks had to deal with quite general issues (as opposed to subject specific questions). Still, some tasks were developed which were especially suited for those teaching young and older children respectively, and for the latter group, some tasks addressed particular subjects (Mathematics and Science/Geography). It should be noted, however, since no references were made to factual or conceptual knowledge in the objectives, such knowledge was also not explicitly addressed in the examination.

Criteria for teacher competency

In sharp opposition to the argument of ”technical-rationality”, where formal instruction is expected to be easily transferred to practice, it has been argued that participation in a community of practice and non-formal learning are the primary routes for learning to become a professional; starting as a peripheral participant and slowly advancing towards a more central position (Lave & Wenger, 1991; Schön, 1983; Wenger, 1998). Following this line of thought, the best way to educate professionals (like teachers) would be to let them participate in the particular community of practice in which they are to become active members. This view is supported by in-service teachers, as they typically claim that the main way of learning to teach is by doing the job (Knight, Tait, & Yorke, 2006; Metcalf, Ronen Hammer, & Kahlich, 1996).

There are, however, some potential drawbacks. For one thing, apprenticeship and non-formal learning are typically more time consuming than formal instruction. This could at least partly be attributed to the fact that novices do not know what is important and what to look for in novel situations. Another, and perhaps more serious point, is that workplace-based training might promote socialization into an unwanted occupational culture and outdated practices (Elliott, 1991). This has been shown to be the case for student teachers, where students have a tendency to demonstrate less sought-after classroom performances after field experiences (Evertson, Hawley, & Zlotnik, 1985). Furthermore, workplace-based training does not aid students to develop skills in reflecting on their practice (Metcalf et al, 1996; Tabachnik, Popkewitz, & Zeichner, 1979).

Acknowledging the potential drawbacks of workplace-based training, however, is not tantamount to arguing that teacher education should be entirely theoretical or entirely campus-based. Instead, it suggests a distinction as to which competencies are best learned (and assessed) in workplace settings and those more properly learned (and assessed) in other settings. This is because there seem to be limitations as to what can be learned through participation, or through more “vicarious means”. Regarding the latter, Elliott (1991) notes that even though professional learning (just like

any other learning) is situated and experiential, it does not have to involve direct participation. Practical situations can also be experienced vicariously, for example by reflecting on case studies and/or discussing different ways to act in relation to simulation exercises. This means that when assessing other competencies than actual teaching performance, the classroom is not necessarily the optimal setting for the assessment to take place. Instead, other settings can offer alternative ways to support the development of specific skills, which is shown by research indicating that different kinds of technology (such as tape recorders in Anderson & Freiberg, 1995; computer simulations in Yeh, 2004; and video in Yerrick, Ross, & Molebash, 2005) can provide effective support for student teachers when analyzing their own, or others', instruction.

On the other hand, when it comes to the tacit knowledge of how to act in the classroom, illustrated below by Berliner's (2004) description of the proficient teacher, this competency is probably best learned by observing experienced teachers and by active participation:

> This holistic recognition of patterns allows the proficient teacher to predict classroom events more precisely. Compared to a novice, they can predict when a student might start to act out, when the class begins to get bored, or when their students are confused or excited. (Berliner, 2004, p. 207)

But even if this knowledge is tacit and presumably difficult to formulate as simple rules, this does not mean that criteria for quality performance cannot (or should not) be formulated. If someone could communicate the criteria of quality performance to the novices, then they would not have to work it out all by themselves. Explicit criteria might in this way help the novice to focus on relevant details in the performance of others, and thus potentially make the field experience more effective, as well as making it possible for students to self-assess their own performance. The central question is thus whether criteria can, or cannot, be articulated in activities characterized by tacit knowledge – such as teaching.

To articulate "tacit knowledge"

Polanyi (1967/1983), who introduced the concept of "tacit knowledge", argued that all human activities, even those that are highly theoretical or scientific, have a tacit dimension. This tacit knowledge, which is grounded in unspoken traditions and experience, provides the frames for how to interpret problems, and how to go about solving them, within a given community of practice.

According to recent theories on learning, all knowledge is tied to human activity, which means that knowledge can never be considered a purely intellectual phenomenon – or vice versa (Säljö, 2005; Wells, 2001). To be regarded as a competent professional, it could therefore be argued that you should not only be able to perform the necessary tasks, but also know *why* you are doing what you are doing. However, this does not necessarily mean that you can articulate the knowledge you possess. Dreyfus and Dreyfus (1986) even argue that an expert can only articulate the rules and theories she learned before she became an expert. This is because an expert is characterized (in their theory) by reacting more or less intuitively in professional situations. An expert works in an automatized and holistic manner, while analyzing and reasoning are characteristics of lower levels in the continuum from novice to expert (Flyvbjerg, 2001; Lindström, 2001).

Säljö (2005), on the other hand, argues that language plays a significant role in *all* human activity, since language provides us with the ability to structure the world and focus on relevant features. As an example of how to articulate "tacit knowledge", he uses the linguistic formulation of "Archimedes' principle". When Archimedes articulated this principle, he put something in words that had been known by many people for several years, but only intuitively. Archimedes' contribution was therefore to make explicit the "tacit knowledge" already known by the merchants. Säljö claims, however, that this difference is paramount from a learning perspective. This is because, once the principle has been articulated, the knowledge can be transferred from the original context and made available for discussion and reflection (see also Svingby, 2005).

Returning to the "tacit knowledge" of teachers, this means that if criteria for high-quality teaching can be articulated, these criteria

can also be transferred from the context of every-day practice, and made available for discussion and reflection. This would facilitate not only to teach students how to become competent teachers, but also to assess this competency.

Formulating criteria

On the topic of assessment, there are some empirical studies indicating that it is possible to articulate the "tacit knowledge" of assessors. Prominent in this field is the work by Lindström (1999, 2001, 2006, 2008). In one of his studies, authentic criteria (i.e. in the sense that they were actually used by those active in the community of practice) were articulated for handicraft work, by using a method known as "repertory grids" (see Lindström, 2001, 2008). According to Lindström, this method is well suited to extract and articulate parts of the "tacit knowledge" within a community of practice, and in this particular study, five categories of assessment criteria were formulated, which were shown to be used by experts in the field of handicraft.

In another study, Lindström (1999, 2006) created a scoring rubric for assessing creativity. As opposed to the inductive design in the study on handicraft criteria, the criteria in the rubric for creativity were formulated on the basis of previous research and current discourse (for example as expressed in journals), and then tested by art teachers.

In the research on teacher competency in this dissertation, criteria have been formulated in a deductive manner, similar to Lindströms criteria for creativity. Starting from course objectives, criteria and standards were formulated in collaboration with professional educators for two parts of teacher competency: (1) the ability to analyze classroom situations, and (2) to self-assess their analysis by comparing own performance to that of a professional.

These two parts of teacher competency clearly only comprise a fraction of the overall teacher competency, and there is much more that students need to be able to do, before they can be considered competent teachers. For instance, teachers must be able to interpret the curriculum, plan instruction that is in line with course objectives, assess students' performance, etc. But by creating rubrics and

standards, a definition of the "competent teacher" can eventually be formulated in terms of teacher performance.

There are already some studies that, together with the rubric in this dissertation, can be used as starting points for such a definition. For example, Osana and Seymore (2004) have developed a rubric for critical thinking skills. In their study, student teachers improved their skills in using research findings when making decisions about complex problems. In another study, Schacter and Thum (2004) created a rubric and formulated teaching standards on the basis of extensive research on teacher quality. These authors demonstrate that teaching performance, as defined by their rubric, actually predicts students' academic progress quite well. Furthermore, Giertz (2003) has created a comprehensive rubric for assessing teaching skills for teachers in academia.

The "Interactive Examination" for dental students

As was argued previously, an assessment methodology that permits to assess students' self-assessment skills, and in this way helps them in developing these skills, would make a great contribution to teacher education. In fact, such a methodology has been developed by Mattheos, Nattestad, Falk Nilsson, and Attström (2004b) for dental students, and it is called the "Interactive examination". In the "Interactive examination" for dental students, self-assessment skills are assessed in parallel to subject-specific skills. Mattheos et al. (2004b) have thus included the assessment of self-assessment skills in a regular examination in an authentic manner, which is in line with the conclusions on self-assessment made earlier. Students' self-assessment skills are assessed by both quantitative (formative) and qualitative (formative and summative) methods, where the center of attention is not merely the quantitative agreement between students and teacher. Furthermore, modern information- and communication technology is used in order to facilitate the assessment – an important point that was raised previously, since one of the problems with introducing performance assessments is that these assessments are often more time consuming and costly than paper-and-pencil tests.

The "Interactive examination" consists of six stages, described below and in Figure 2. The description, which is based on the presentation made in Jonsson, Mattheos, Svingby, and Attström (2007b), is provided in order to show the general outline of the methodology as it was originally designed.

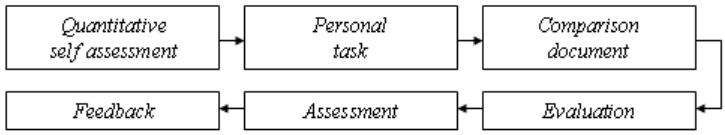


Figure 2. A graphic representation of the six stages in the "Interactive examination".

Quantitative self-assessment. The students assess their competency through a number of Likert-scale questions graded from 1 to 6. This self-assessment is then compared to a corresponding assessment made by the instructor. The purpose of the comparison is to highlight possible differences, which can serve as a basis for reflection or discussion. The purpose is not to assess students' self-assessment skills as such.

Personal task. When the students have finished the initial self-assessment, they get a personal task in the form of an authentic, professional problem. The students must come up with a solution to the problem, and their answers are submitted as written documents.

Comparison document. As soon as the students submit their answers to the personal task, they can access a professional solution to the same problem. This professional solution is not seen as a "right answer", but rather as an answer from an experienced colleague; an answer which is well supported by references to practice and research. With this professional answer as a starting point, the students compile a "comparison document" where they identify differences between their own response and that of the professional. The students are also expected to discuss these differences in the documents and try to state own learning needs. This comparison is

thus a qualitative form of self-assessment, which, in contrast to the quantitative self-assessment, is used for summative examination purposes as well.

Evaluation. After the examination, but before they get feedback on their results, the students complete an evaluation questionnaire. In this questionnaire they are asked about things like software navigation and task difficulty, but also about their perceptions of learning.

Summative assessment. Both the personal task and the comparison document are assessed for summative purposes. In the personal task students' subject-specific skills are assessed, while their skills in reflecting on their choices, in identifying weaknesses, and in stating own learning needs are assessed through the comparison document.

Feedback. About a month after the examination, individual feedback is sent to all students, commenting on the quantitative self-assessment, the personal task, and the comparison document. If a student has not reached an acceptable level of quality on any one of the tasks used for summative purposes, she will be assigned additional tasks until an acceptable level is reached.

In the research by Mattheos et al. (2004b), it was noted that the methodology is appreciated by the students. For example, the possibility to reflect on their self-assessment and to identify own learning needs were stressed by the students as strong points of the examination.

Mattheos et al. (2004b) show that it is quite possible to design an examination for assessing self-assessment skills in an authentic way using ICT. What might be considered problematic, however, is that the assessment of the comparison document is done without any clear criteria. This means that there might be problems of credibility in the scoring procedure, and that students do not know what is expected of them. Furthermore, it might be difficult and time consuming to give feedback, which was also noted by the authors. These problems could possibly be remedied by constructing a rubric for the comparison document. This was done in the version of the "Interactive examination" developed for student teachers.

The "Interactive examination" for student teachers

In the same way as dentists are expected to adapt to new technologies and new ways to diagnose and treat patients, professional teachers should be able to continuously evaluate their own instruction and change it accordingly (Yeh, 2004). Thus, the "Interactive examination" methodology, with its focus on self-assessment of professional competency in an authentic context, has great relevance to teacher education as well. The innovative design of the "Interactive examination" for dental students was therefore adapted for teacher education, and implemented in that context.

The two versions of the "Interactive examination" have a similar structure (see Figure 2). The main differences are in the design of the personal task and the professional document. These variations stem from differences in the respective professions. Other differences between the two versions, such as the fact that the student teachers are assessed with the support of a rubric, originate in the assumptions on how to make the assessment credible, as well as support learning, a feature which was incorporated into the design of the "Interactive examination" for student teachers only. The differences between the "dentist" and "teacher" versions are presented in more detail below.

The personal task

As opposed to clinical problems in the field of medicine, which were used as cases for the dental students' personal task in the work by Mattheos, Nattestad, Christersson, Jansson, and Attström (2004a), teachers' work is often less well defined. Teacher competency involves the handling a wide variety of situations. These range from providing appropriate conditions for student learning, to attending to an individual student's social and psychological difficulties, but also involve assessing student knowledge, arranging meetings with students, parents, and/or colleagues, etc. The personal task in the "Interactive examination" for student teachers had to reflect a similar complexity, for the examination to provide valid data about student performance. Therefore the tasks did not focus on details or well defined problems. Instead, they were (more or less) open for interpretation, so that the students themselves had

to choose what was important and identify one or more problems to be solved.

The tasks chosen for the "Interactive examination" for student teachers were classroom cases, simulated through digital-video recordings. Inspiration for the design came in part from a computer-based patient-simulation system, which was developed in order to facilitate the use of realistic and interactive virtual patients for educational purposes. Evaluations of that system have shown that students found the cases to be engaging, realistic, instructive, and fun (Zary, Johnson, Boberg, & Fors, 2006). The classroom situations included in the "Interactive examination" for student teachers were developed in cooperation with representatives from the School of Teacher Education, as well as with in-service teachers. The intention was to display genuine and problematic situations which had no correct or obvious solution. The situations were then filmed in schools, with teachers and students acting.

Research has shown that student teachers might need guidance in order to structure their learning when working with simulated situations. For example, in a study by van den Berg (2001), it was noted that students' reflections remained at quite a superficial level, dealing more with "classroom management" than with more general and theoretically grounded issues. Similar results, where beginning and novice teachers provided simpler and more descriptive comments on teaching events than "expert teachers" did, are reported in a number of studies (e.g. Berliner, 1986; Carter, Cushing, Sabers, Stein, & Berliner, 1988; Lin, 1999). Thus, in order to provide a "scaffolding structure" for how to analyze the cases, a review was conducted of empirical research on student teachers' reflections on classroom situations with the aid of technology. This research indicates that technology can provide effective support for student teachers' learning when analyzing their own, or others', instruction (e.g. Yeh, 2004; Yerrick et al., 2005). For instance, in a study by Metcalf et al. (1996), student teachers' "reflective ability" was measured. Reflective ability could be defined as the ability to describe and analyze complex teaching episodes. Students' written descriptions and analyses of cases were analyzed as to the level of reflection they displayed. The levels of reflection used for categorization were: (1) low level "factual discourse", where the student

could provide factual information about pedagogical actions; (2) medium level "prudential discourse", where the student could give suggestions for improvement; and high level "justificatory discourse", where the student could give reasons or rationales for pedagogical actions.

A group of students in the study by Metcalf et al. (1996) were exposed to a series of campus-based activities, which included role play and giving short lessons which were videotaped. Furthermore, the students watched simulated classroom situations on video. With these situations as a starting point they had to provide explanations, suggest possible solutions, and propose potential consequences of the situations. Every activity was analyzed and discussed by the students in groups.

The "reflective ability" of these students was then compared to a group of students who had been exposed to regular field-based education. Results from this comparison showed that the group of students exposed to campus-based activities had significantly improved their skills in identifying critical events in complex, pedagogical situations. They could also give more advanced explanations to these situations and were more inclined to give rationales for different actions taken. Even skills in carrying out meaningful lessons, as measured in this study, were improved by this group, whereas the skills of the control group had not changed.

The results from Metcalf et al. suggest that, by working systematically with simulated situations (such as video sequences and role play), student teachers can (1) improve their skills in identifying critical events in complex situations, (2) give more advanced explanations to these situations, (3) become more inclined to give rationales for actions taken, as well as (4) improve their performance in giving lessons. Therefore the concept of "reflective ability" proposed by Metcalf et al. was used as a scaffolding structure in the "Interactive examination". With the movies as a starting point, the students responded to three global questions, which were the same for all movies:

1. Describe the situation without prejudice (Observation).
2. State a problem and analyze the situation displayed (Analysis).
3. Formulate which actions should be taken, considering the needs of all those involved (Taking action).

These global questions were not broken down further, since it might have affected students' perceptions of complexity and authenticity negatively, if they had to give answers to a set of well defined questions, as opposed to analyzing the situation more freely. This problem may have occurred in the study by van den Berg (2001), where the situations were accompanied by a number of tasks, for instance to make connections between the case and the most relevant section in the text-book. These tasks were thought to aid students' reflections, but some students directed their efforts towards solving the task in an uncritical and instrumental way.

Screenshots from the "Interactive examination" for student teachers, and examples of student answers are shown in Appendix A.

The professional document

The main purpose of the professional document is to provide an exemplary answer, in relation to which the students assess their competencies and identify their shortcomings. This is accomplished by comparing their own analyses and solutions to that of the professional.

However, given the fact that there are no correct answers to the situations displayed in the movie sequences in the "Interactive examination" for student teachers, and rarely any recommendations of "best practice" of the kind found in medical professions, the role of the professional document is to analyze the situations as thoroughly, and from as many perspectives as possible. The professional documents thus do not provide a single well grounded analysis and solution to the problem, but rather outline several potential approaches to problem solving.

The rubric

To achieve as high reliability as possible, the rubric constructed was topic-specific and analytic (Appendix B). For each of the glob-

al questions (Observation, Analysis, Taking action) four or five assessment criteria were formulated. This was done by integrating the educational objectives of the course with the questions derived from Metcalf et al. (1996). The harmony of criteria and objectives was imperative to avoid domain-irrelevant variance in the assessment, which was also important when formulating levels of quality (Acceptable and Excellent) for each criterion. Examples of how the objectives have been operationalized into criteria are given in Table 1. Here it can be seen that the skills of documenting and describing situations are assessed through the Observation task. This is done by asking the students to describe the situation displayed, where the qualities of this description are further operationalized into four different criteria (see Appendix B). It can also be seen that the skills in using relevant scientific material, as well as own (and others') experiences as a basis for professional development, is assessed through the Self-assessment task. This is done by asking the students to compare their own answers to that of a professional, and by asking them to argue for their own standpoints with the help of course literature or other relevant sources (the use of research-informed arguments is also assessed through the Analysis and Taking action tasks, as seen in Appendix B).

In order to support student learning, the rubric was designed according to the principles for formative assessment by being task-related and providing quality standards. As discussed in previous sections, however, handing students a rubric does not automatically lead to improvement of performance. Some kind of instructional intervention is also needed, whereby the students become accustomed to the criteria. This is an in-built feature of the "Interactive examination", since both quantitative and qualitative self-assessments for formative purposes are part of the methodology. Additionally, as opposed to the "Interactive examination" for dental students, assessment criteria were formulated for the qualitative self-assessment (the comparison document) as well, resulting in a rubric with a total of 16 criteria. The rubric was distributed to the students approximately three weeks before the examination, so that they could read and discuss the criteria with peers and instructors.

Table 1. Examples of how course objectives were operationalized in the rubric.

The student teacher should be able to:	*Task*	*Examples from the rubric*
- document, describe, and reflect on students' situations.	Observation	The description focuses on relevant details [in the situation displayed].
- use relevant scientific material as well as own (and others') experiences as a basis for professional development.	Self-assessment	The comparison [with professional document] argues for own standpoints with the help of course literature or other relevant sources.

Comparison of quantitative self-assessment

In the quantitative self-assessment, the dental students assessed themselves on a scale from 1 to 6, and the clinical instructor assessed the students according to the same questions and the same scale. These assessments were then compared and differences were noted. The student teachers also assessed themselves on a scale from 1 to 6, but since there were no assessments available from mentors (corresponding to the clinical instructors), comparisons were instead made with the actual results of the examination (i.e. the self-assessment was restricted to the examination context). The comparison of self-assessment and examination results was made possible by reformulating the assessment criteria as self-assessment questions, thus making each question correspond to a criterion in the rubric (Appendix A).

Furthermore, the results on the examination were reported to the students for each individual criterion on a scale from 1 to 6, which facilitated the comparison with the self-assessment. The assessment was performed in relation to three levels of quality (Fail, Acceptable, Excellent), which were transformed into marks (0, 1, 2). As each student analyzed three different movie sequences, she received a score somewhere from 0 to 6 for each criterion. For research

purposes, an overall score was computed from the sub-scores for the separate criteria, for example in order to estimate the inter-rater reliability. However, no overall scores were communicated to the students, since no other grades than Pass or Fail were used in the course in question.

Developmental changes in the "Interactive examination" for student teachers

The "Interactive examination" for student teachers was piloted in 2004, and a thorough analysis was subsequently made of students' answers and results. This analysis resulted in three major changes, which were implemented in the 2005 version of the examination. These will be described in more detail below. No further changes were made between the 2005 and 2006 versions of the examination.

After the pilot version of the examination, frequency analyses were carried out for students' results on each criterion in the rubric. The rationale for these analyses was that any criteria which a majority of the students did not meet, or only barely met, were indistinctly expressed, and not that these criteria necessarily were more difficult than the others, or that the students who succeeded were more gifted. This analysis led to a reformulation of some of the criteria in the rubric. For example, the first criterion, assessing students' observational skills, was originally expressed "The description is not prejudiced". This was later revised by adding the following sentence to the "Acceptable" level: "The description may contain assumptions that are not shown in the situation displayed", thus more clearly expressing what is meant by this criterion.

Another part of the analysis focused on what levels in the rubric were not, or seldom, used by the assessor, since this could indicate that he had encountered difficulties discriminating between the different levels. In those cases where it was not possible to further clarify the difference between the levels, or where the difficulties in discriminating between the levels most likely were due to the design (i.e. that the tasks or the methodology did not give the stu-

dents the opportunity to show their proficiency), the levels were merged so that no discrimination was needed. In one case, a criterion was removed from the rubric, which explains why the pilot version of the rubric contained 16 criteria, while the 2005 and 2006 versions only had 15. Since the questions in the initial quantitative self-assessment are equivalents of the criteria in the rubric, these questions were also changed accordingly

After piloting the examination, a qualitative analysis was made of the comparison documents written by the students. The results from this analysis were compared to a similar examination for dental students (Jonsson et al., 2007b), which showed that the dental students seemed to have a somewhat different attitude towards their experienced colleagues than the student teachers did. While the dental students saw more authority in the experienced dentist, the student teachers to a greater extent regarded their own answers as being as good as the professional's, or even better. There are many possible explanations for these results, but one likely hypothesis stems from the difference in the nature of knowledge used in the two professions. While dentists rely mainly on scientific and evidence-based knowledge, more psycho-socially oriented pedagogical skills are needed in the teaching profession. Another hypothesis is that the professional solutions were not perceived as sufficiently professional by the student teachers. Therefore these documents were revised for the 2005 examination, in order to present a more thorough and systematic analysis of each movie sequence.

In the analysis of the comparison documents described above, a categorization was made of students' comparisons, for example which kinds of differences the students identified between their own answer and that of the professional (Jonsson et al., 2007b). This categorization made it possible to clarify the "self-assessment criteria" in the rubric. For example, one of the criteria was originally expressed as "The comparison identifies differences between own and the other's interpretation of the situation displayed". But since some differences are more interesting than others (it is for instance quite uninteresting to notice that the number of words differ, while differences related to subject matter could be regarded as more important), this criterion was separated into more distinct levels, depending on the kinds of differences identified.

Besides the changes in the rubric and the expert documents, student answers from the 2004 cohort were used to compile a document with answers assessed and commented upon in relation to the rubric (Appendix C). Aiming to help the students to interpret the criteria, this document was distributed to the students before the examinations in 2005 and 2006. All the student answers in the document were taken from one specific movie, which was then removed from the pool of movies used for the actual examination.

To summarize, between the 2004 and 2005 versions of the "Interactive examination", the following changes were made: (1) a number of criteria in the rubric were expressed more distinctly; (2) the quality levels for three criteria were merged, so that no discrimination was needed; (3) one criterion was removed; (4) questions in the initial quantitative self-assessment corresponding to the criteria changed in the rubric were also changed accordingly; and (5) the thoroughness and professional appearance of the professional solutions were enhanced. Furthermore, the students in cohorts 2005 and 2006 were able to (6) access a document with assessed student answers ("exemplars").

Research methodology

Research questions

To investigate whether the competencies aimed for in the "Interactive examination" for student teachers can be assessed in a credible manner, and whether the examination methodology supports student learning, the following research questions were explored[11]:

1. Does the methodology "work"? (Study II)
2. Is the methodology valid for both summative and formative purposes? (Study III)
3. Does the use of transparency improve student performance? (Study IV)

[11] See each article for more detailed information on these questions.

Sample

The examinations providing data for this dissertation were carried out in the fall of 2004, 2005, and 2006 respectively, all with a cohort of first year student teachers in Science, Geography, and Mathematics (n = 432; 170, 154, and 138 respectively). The students had not been exposed to the "Interactive examination" previously.

Research data and analyses

In accordance with case-study methodology, several different sources of data have been used, such as demographic data, students' (quantitative) self-assessments, students' examination scores, a student evaluation questionnaire, and students' comparisons with professional documents. Sources of data and analyses are described only briefly here; more thorough descriptions can be found in the articles. An overview is provided in Table 2.

Demographic data. Demographic data were collected through a web-based questionnaire, including gender, parents' education, ethnicity, computer experience, subject major, instructor, and specialization towards teaching higher or lower levels. To explore the effects of these variables, regression analysis was used.

Quantitative self-assessment. In the self-assessment stage of the examination, the students estimated their own competency through Likert-scale questions. This estimation was then compared to the actual examination results, using a two-tailed Wilcoxon signed-rank test, revealing if the students assessed themselves to be better, worse, or in agreement with their results. This provided a rough estimate of the students' self-assessment skills, referred to as their "self-assessment pattern".

Students' examination scores. The students were assessed on their analyses of critical classroom situations, as well as on their comparison with the professional analyses of the same situations. Assessor consistency was estimated by using Cronbach's alpha, while an inter-rater reliability analysis was performed using exact agreement (in percent) and rank correlation (Spearman's rho) at the criterion level, along with Pearson's correlation for the overall score.

Table 2. An overview of data collected, and analyses performed, in relation to the different studies on the "Interactive examination".

Data	*Analyses*	*Study*
Demographic data	Effects of background factors on examination score (Regression analysis)	Study II-IV
Quantitative self-assessment	Students' self-assessment patterns (Wilcoxon's signed-rank test)	Study II
Students' examination scores	Assessor consistency (Cronbach's alpha)	Study III
	Variance components for students Dependability coefficients (Generalizability theory)	Study III
	Inter-rater reliability analysis (Exact agreement, Spearman's rho, Pearson's *r*)	Study III
	Differences in mean score between cohorts (ANOVA, Effect size)	Study IV
Comparison documents	Comparison with professional (Qualitative analysis)	Study II
Student evaluation questionnaire	Students' perceptions of learning, authenticity, etc. (Medians, Non-parametric tests)	Study II-III
	Rubric use (ANOVA, Effect size)	Study IV

Generalizability theory was used to estimate the magnitude of variability due to students, and dependability coefficients were computed for different numbers of movie sequences in order to see how many movies would be needed in order to consider the scores generalizable.

Comparison documents. The students compared their own answers with the professional documents, and identified differences. They were also expected to comment on the differences, as well as to identify own weaknesses and learning needs. A qualitative analysis of students' answers to the comparison task was performed

The student evaluation questionnaire included a total of 20 questions. In the questionnaire, the students were asked about their perception of different features of the "Interactive examination", for example their perception of alignment between assessment and instruction, and whether the scoring criteria had been comprehensible to them. Several of the questions were measured on Likert scales from 1 to 9, and median values were computed. This particular scale was used for easy comparison with the "Interactive examination" for dental students (see Mattheos et al., 2004b).

Methodological limitations

Pre-requisites for the "Interactive examination" for student teachers were that it should be used for both summative and formative purposes, and that it should be implemented into an existing course in teacher education. The latter condition means that the research performed in relation to the methodology can be argued to have high "ecological validity" (see Black & Wiliam, 1998a). A particular strength of such research in assessment is that the students have a strong motivation for doing well on the examination, since it has real-life consequences for them. On the other hand, there are also a number of drawbacks in optimizing an assessment for student learning, since these settings do not always provide the most favorable conditions for research.

An important example of such a weakness is that analyses made on students' use of rubrics were based on students' own reports in the evaluation questionnaire. Experimental conditions were not imposed, and the main reason for this was the ambition to give all students equal and as optimal conditions as possible. Also, it could

well be considered unethical to provide only some students with a scoring rubric in an examination used for summative purposes, since they would not be "competing"[12] on the same terms. Since the students were not selected at random, potential differences between students using and not using the rubric should be interpreted with care. For example, it might be that the more ambitious students were also the ones using the rubric, and perhaps these students would have performed better even without this support.

Another problem occurs when generalizability theory is used. As suggested by Linn et al. (1991), the magnitude of variability due to the sampling of tasks should be included in a generalizability analysis. In the "Interactive examination", however, each student could choose three movie sequences from a pool of nine, resulting in what is known as a "nested design". Unlike the "crossed design", the nested design has no separate term for the item effect, which is instead part of the residual. Since students had chosen different movies, the movie effect could not be estimated independently of the person-by-item interaction. The nested design thus suffers from a disadvantage as opposed to a crossed design generalizability study (Shavelson & Webb, 1991), and only variance components for students, as well as the movie effect confounded with the residual, were computed.

The fact that the research is based on one case only means that the knowledge produced is contextualized and temporary, which could possibly be classified as a weakness in research design. This, however, depends on the view taken on generalizability issues. Proponents of the hermeneutic tradition have for a long time made a clear distinction between natural and social sciences (e.g. Dilthey, 1900/1996), as to what kinds of theories can be formulated, and consequently what claims of generalizability can be made. Whereas the aim of natural sciences is to find general patterns, or "laws", which can be assembled into universal, scientific theories, such predictive theories should not be expected in the study of human interaction. As opposed to the natural sciences, it is not possible to formulate universal laws on how people act, since every individual and every situation is in some sense unique and guided by human

[12] The grading procedure is not competetive in itself, since grading is not cohort- or norm referenced. Still, there might be a competition regarding qualifications for future courses or job opportunities.

intentions. The social sciences thus have to start from specific cases, without being able to generalize the results in the same sense as in the natural sciences (Flyvbjerg, 2001). As a consequence, the question of so called "analytic", or "qualitative", generalizability is often discussed in relation to case studies. In principle, "analytic generalizability" means that the level of generalizability is estimated by the reader/user herself, on the basis of the evidence provided by the researcher (Stake, 1994). The researcher thus has to present her/his research in a way which makes it possible for the reader to assess to what extent the findings generalize across different circumstances, for example from higher education to schools or from laboratory settings to classroom contexts. As discussed previously, the first parts of this dissertation were eclectic in this regard, citing for example a wide range of research on rubrics. Nevertheless since the problem of interest is the use of performance assessment in teacher education, the empirical studies have been carried out in a teacher education context, and are implemented in natural settings.

STUDY II-IV: THE "INTERACTIVE EXAMINATION"

The "Interactive examination" was studied during three consecutive years: 2004, 2005, and 2006. In the following, brief summaries of the three studies are given, and the main conclusions in relation to the research questions (see p. 84) are presented. The studies are presented in more detail in each article.

Study II: Does the "Interactive examination" for student teachers work?

In study II (Jonsson et al., 2007b), the "Interactive examination" was performed with dental students and student teachers in parallel. One of the main purposes of the article was to present the methodology as it was operationalized in the two environments. Another focus was on how the methodology "worked". That the methodology "works" can in this context imply practical issues (i.e. if the interface is easy to use, if the sound quality and movie resolution were sufficient, etc.), but can also refer to issues of validity. In connection to validity, it was investigated whether students performed as intended, and how they perceived the examination (sometimes referred to as "face validity"; see Brown, Bull, & Pendlebury, 1997).

Results and conclusions

Results from the study show that the students appreciated the methodology, and also that it works in a comparable way in both en-

vironments. For instance, it can be seen that both dental students and student teachers responded to the tasks in a similar manner. However, some interesting differences were also discovered[13]:

1. The self-assessment pattern (i.e. if the students assessed themselves higher, lower, or in agreement with the clinical instructor or the assessor) was quite different in the two groups. Among the dental students, there were students assessing themselves both significantly higher and significantly lower than the clinical instructor (38 % and 24 % respectively), which can be compared to the fact that the majority of student teachers assessed themselves significantly higher than the assessor (72 %).
2. Students' attitudes towards the "experienced colleague" (in the professional document) were also quite different in the two groups. Although the task was to identify differences between their own answers and those of the professional, most of the student teachers chose to highlight similarities. Furthermore, several student teachers argued against the professional, sometimes arguing that they had produced an answer that was qualitatively better than the professional's. A similar tendency could not be seen in the answers provided by the dental students.

The conclusions from Study II are that, by investigating how the methodology was used and perceived in the two different institutions, a rough estimation of the validity could be made (i.e. if the methodology "works"). Even though institutional differences exist, displayed for instance in the different ways of handling the comparison task by student teachers and dental students, the overall applicability of the methodology is similar in both centers, and the students respond to it in an analogous manner. This indicates that the "Interactive examination" "works" as intended.

[13] Both of these findings have been investigated further by making changes in the methodology (e.g. by changing the appearance of the professional documents), but since these analyses are not part of the papers in the dissertation, the results are not presented here.

Study III: Is the "Interactive examination" for student teachers valid for its summative and formative purposes?

In line with the problems explored in this dissertation, it is of interest to investigate whether the examination is credible (summative purpose), and whether it can support student learning (formative purpose). Study III (Jonsson, Baartman, & Lennung, 2007a) therefore investigates whether the "Interactive examination" can be considered a valid assessment methodology for both summative and formative purposes. The distinction between summative and formative is made, since – as Messick (1996) points out – validity is not a property of the instrument, but differs depending on how the scores are interpreted and used, something that clearly differs for the summative and formative use of the same instrument.

The validation was carried out by the support of a framework for quality estimation of competency-assessment programs, called the "Wheel of competency assessment" (Baartman, Bastiaens, Kirschner, & Van der Vleuten, 2006) (Figure 3). This framework was developed from the notion that in order to evaluate the quality of competency assessments, criteria to determine their quality are needed. Since competency assessments might consist of both traditional and new forms of assessment, criteria used to evaluate the quality of competency assessments should be derived from both psychometrics and edumetrics[14] (Baartman et al., 2008). The "Wheel of competency assessment" was therefore developed by taking Messick's theory of construct validity as a starting point, together with contributions from Frederiksen and Collins (1989), Linn et al. (1991), as well as other contributions on this topic. The criteria were then validated by experts on assessment (Baartman et al., op. cit.), but also by teachers (Baartman, Bastiaens, Kirschner, & Van der Vleuten, 2007).

[14] Educational measurements. See e.g. Gielen et al. (2003).

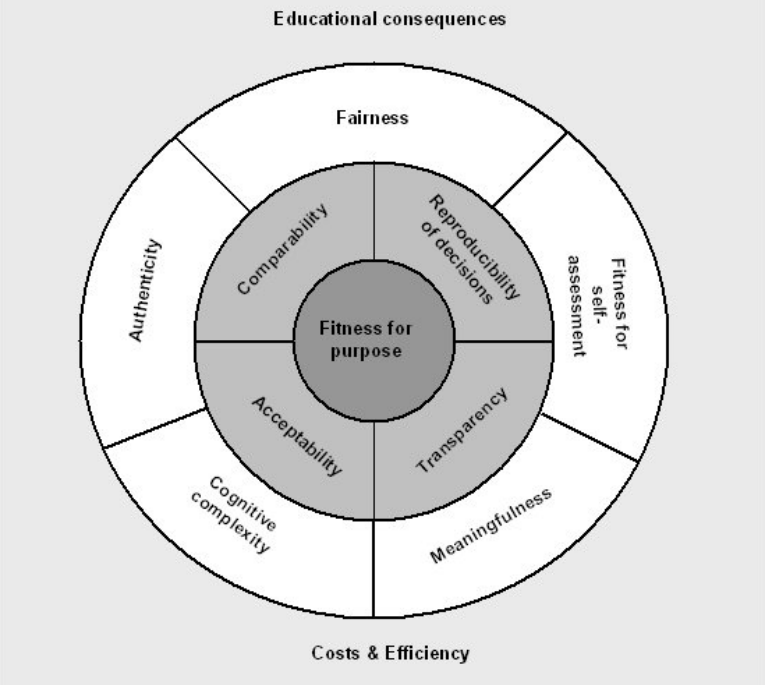


Figure 3. The "Wheel of competency assessment". Adapted from Baartman et al. (2006).

In the "Wheel of competency assessment" the quality criteria are displayed in concentric circles. Fitness for Purpose forms the basic quality criterion for all competency assessments, and is related to the idea of "constructive alignment" between learning, instruction, and assessment (Biggs, 1996). The inner layer consists of Comparability, Reproducibility of decisions, Acceptability, and Transparency, which are seen as the more basic quality criteria. The outer layer represents more newly developed quality criteria: Fairness, Authenticity, Cognitive complexity, Meaningfulness, and Fitness for self-assessment. The wheel is placed in a broader educational context through the criteria of Costs & Efficiency and Educational consequences.

There are several reasons for using this particular framework, the major reasons being that it is more comprehensive than many other frameworks suggested, and that it contains criteria for formative as well as summative purposes. Even though Messick's theory of construct validity certainly means a broadening of the concept of validity, as compared to traditional forms of validity, there are nevertheless problems with using this validity concept

when dealing with formative assessments. One such problem is transparency, which, as described previously, is not represented in Messick's framework. The same is true for the use of self-assessment and/or feedback in the assessment, both of which are seen as important requirements in formative assessments (Black & Wiliam, 1998b; Frederiksen & Collins, 1989; Wiggins, 1998).

There are, however, also some problems associated with using the "Wheel of competency assessment". First of all, although presented in concentric circles, there is no compelling theoretical rationale for this arrangement, or for the relationship between the different criteria. As a consequence, the concentric arrangement is not used in the current study. Instead the criteria are re-arranged, so that they are grouped on the basis of their importance for the validation of summative and formative purposes, respectively (or as important for all assessments).

Another problem is that the framework only contains criteria, but no standards. Furthermore, there are no suggestions or rules on how to operationalize the criteria. So when validating the "Interactive examination" according to the criteria in the framework, operationalizations and standards had to be formulated. For the summative purpose, standards could be found in the literature, for instance what is to be considered an acceptable level of inter-rater agreement, and these standards have guided the validation process (still, it should be emphasized that other operationalizations of the criteria could naturally be made). For assessments with formative purposes, the situation is quite different, since it is much more difficult to find standards for what should be required for formative assessments, or for combinations of formative and summative assessments.

To summarize, the usefulness of the "Wheel of competency assessment" lies mainly in its comprehensiveness, since it provides a thoroughly validated list of important aspects to be considered, representing both psychometric and edumetric concerns. The concentric arrangement, however, has not been used in the current study. Instead, a re-arrangement of the criteria was made to fit the specific needs when validating the "Interactive examination" for its summative as well as its formative purpose.

The lack of standards and guidelines on how to operationalize the criteria was most noticeable when validating the formative purpose, and in the research presented here students' perceptions of the assessment were used as the main data source for this part of the validation process. The reason for using students' perceptions is based on the findings by Entwistle and Ramsden (1983) and others, showing that students' approaches to learning are greatly affected by assessment demands and that it is the context as perceived by the students which affects students' learning strategies. As has been argued, it is thus important to make a distinction between the context as seen from the outside, and how it is perceived by the students (Entwistle & Peterson, 2004). However, since the criteria investigating students' perceptions lack any external standards, these estimates can only be related to each other within the specific context.

The instrument used for data collection was the student evaluation questionnaire. Since the students need to be aware of what is expected of them, if they are to adapt their learning strategies to the assessment requirements, the questionnaire used in the evaluation must not be designed as a psychological test for measuring latent cognitive variables. Instead, as the students are seen as responsible adults, able to understand the purpose and rationale for the examination, they are asked their opinion about certain aspects of the examination.

Results and conclusions

Results from the validation process indicate that the "Interactive examination" is valid for its summative purpose. Although further improvements could be made in relation to the same criteria, there is sufficient evidence of compliance with the criteria important for summative assessments ("Comparability", "Reproducibility of decisions", and "Fairness") in the framework.

The results from the student evaluation questionnaire show how the students perceived the different aspects of the examination used for validating the formative purpose. For instance, the students perceived the purpose of the examination to be very clear. Furthermore, to a great extent, the students thought that the examination was well aligned with the instruction; that it was a valuable

mode of examination; that the examination was authentic; and that the examination was helpful in preparing them for their future work. Other interesting results are that it was not particularly clear how to get a pass grade or how to interpret the criteria; that while time was perceived as a relatively large problem, language was not; and that the examination had not affected their ways of studying, but might have some potential in doing so.

On the basis of these results, it is argued that the ”Interactive examination” is a valid assessment methodology also for its formative purpose, although improvements can be made in relation to some criteria (“Transparency” and “Fairness”). With regard to the criterion “Educational consequences”, it might be further investigated whether the students actually learn through this specific methodology.

Study IV: Does the use of transparency improve student performance?

The starting point of Study IV (Jonsson, 2007) is Frederiksen and Collins’ (1989) influential article ”A systems approach to educational testing”, where they put forth the concept of “transparency” as an important aspect of assessment validity. According to these authors, the terms of the assessment must be clear to the students, otherwise it holds no potential for motivating and directing learning.

In Study IV, the effects of increased transparency in the “Interactive examination” are investigated through the use of:

1. *Self-assessment criteria*, which make explicit how to self-assess in the specific context, and in this way “scaffold” the process of self-assessment;
2. *A scoring rubric*, which combines assessment criteria with different levels of quality performance, making the expectations of the assessment explicit;
3. *Exemplars* (i.e. a document with assessed student answers), which (in combination with rubrics) is thought to enhance the consistency in scoring, as well as providing examples of both

low- and high-quality performance.

The study comprises three consecutive examinations carried out in 2004, 2005, and 2006, all with a cohort of first year student teachers (n = 462; 170, 154, and 138 respectively). Between the 2004 and 2005 versions of the ”Interactive examination”, a number of changes were made to increase the level of transparency, such as expressing some criteria in the rubric more distinctly and giving the students access to a set of exemplars.

Results and conclusions

Results show that there were large differences in the quality of students' answers between the 2004 and 2005 cohorts, when changes in the examination were implemented, and as a consequence the overall score increased with over 60 percent (effect size = 3.21, $p<.001$). The scores for the individual tasks also increased markedly; most notably the Analysis task, where the scores more than doubled from a mean score of 2.00 in 2004 to a mean score of 4.49 in 2005 (effect size = 4.25, $p<.001$). On the other hand, the comparison between 2005 and 2006 (when no further changes were made), does not show a corresponding difference. There is a slight increase for some of the tasks, but they are typically quite small (9 percent at the most), and not all are statistically significant. Furthermore, in one of the tasks, there is a decrease in score from 2005 to 2006.

These results show that by making the assessment more transparent, students' performances could be greatly improved. This study thereby corroborates earlier findings, where the use of rubrics or exemplars has been shown to aid in improving student performance.

DISCUSSION

This dissertation commenced by noting that teacher education is a profession-directed education, where students are expected to become competent professionals, and that aiming for competency means that there is a need for assessment methodologies which assess whether students can *act knowledgeably* in relevant situations. Furthermore, due to the so-called "backwash effect", such assessments should not be limited to only measuring students' acquisition of the competencies aimed for, but also be used to support the development of these competencies.

Yet another issue was introduced, and that was how the students can be prepared for life-long learning and "inoculated" with a continuing ambition to improve their work. It was argued, again with reference to the strong effect of assessment on student learning, that students' skills in reflecting on their performance must not only be taught but also be assessed.

The aim stated for the dissertation was therefore to explore how teacher competency (including self-assessment skills) can be assessed in an authentic manner, and how the assessment can support student learning, while acknowledging the importance of credibility and trustworthiness in the assessment.

In order to explore these issues, a literature review was performed, investigating whether the use of scoring rubrics can aid in supporting credible assessment of complex performance, and at the same time support student learning of such complex performance. The conclusions from this review were then implemented into the design of the "Interactive examination" for student teachers. Through this examination methodology, it was investigated wheth-

er the competencies aimed for in the "Interactive examination" for student teachers could be assessed in a credible manner, and whether the examination methodology supported student learning. From these investigations, it was concluded that the examination is valid for its summative as well as its formative purpose, and that the "Interactive examination" can therefore be used as a basis for grading, while at the same time supporting student learning of the competencies aimed for. This dissertation thus gives an illustration of how formative and summative purposes might co-exist within the boundaries of the same (educative) assessment. Furthermore, the results show that, by making the assessment more transparent, students' performances could be greatly improved. This study thereby corroborates earlier findings, where the use of rubrics or exemplars has been shown to aid in improving student performance.

However, the research in this dissertation has a number of unique features and limitations that need to be taken into account. Several critical questions could be raised in relation to the conclusions above, and some of these questions will be discussed in this final chapter of the dissertation. The discussion then ends with suggestions for future research and implications for practice.

Assessing teacher competency

One of the aims for this dissertation was to explore how teacher competency can be assessed in an authentic manner. But how can it be claimed that it is in fact teacher competency that is being assessed, when the assessment is not performed in authentic settings?

In answering this question, it should be noted that to claim that the "Interactive examination" assesses teacher competency, is to make *an interpretation of what the scores from the examination represent*. The validity of this interpretation depends, in turn, on the credibility of the inferences involved in the interpretation. Kane et al. (1999) describe three different levels of such inferences that have to be credible in order to claim validity of a performance assessment as a whole. First, students' performances are observed and scores are awarded in relation to the criteria. Second, the scores are generalized to a domain of assessment performances

beyond the specific set of tasks in the examination, and third, the score is extrapolated to the target domain. To decide whether it is really teacher competency that is being assessed through the "Interactive examination" is therefore a question of addressing each of the levels in this "chain of inferences".

Most basically, validity in this matter depends on how the criteria are formulated and applied. As expressed by the structural aspect of Messick's construct validity, the scoring structure has to be in line with domain theory in order to be considered valid, and this aspect should be validated by experts' judgments – just like the content aspect (Miller, 1998). In the "Interactive examination" the criteria were developed from course objectives, and the formulation was informed by research as well as by the judgment of experienced teacher educators, ensuring that the scores awarded for student performance represent performance that is indeed sought for in the target domain. Furthermore, the scoring of different assessors was investigated and compared, to see whether they used the criteria as intended. Since the students were thus acting/performing in relation to relevant situations, and were assessed according to valid criteria, it could be argued that it is indeed competency that was being assessed. But to be "competent" also means to integrate knowledge and attitudes in order to act *knowledgeably* – so how do we know that their actions were based on professional knowledge? Once again, this depends on the formulation of the criteria. The criteria in the rubric must ask the students to reveal the arguments for their actions, so that it can be seen whether they use their professional knowledge or not. A pre-requisite for assessing professional competency is therefore the formulation of criteria that not only assess student performance, but also the basis of their actions (i.e. professional knowledge and attitudes; see Giertz, 2003; Gonczi, 1994). In the "Interactive examination", students were asked to argue for their standpoints by referring to course literature or other relevant sources, which means that the use of professional knowledge as a basis for their decisions is rewarded.

Regarding attitudes, these are not explicitly assessed in the "Interactive examination". This is due, on the one hand to the difficulty in ensuring that students' writings represents their genuine attitudes, and is not only an expression of what they know is consi-

dered appropriate, and on the other, to the ethical issue involved in assessing personal values. It could therefore be argued that, in the case of attitudes, only performance should be assessed for summative purposes, and as long as performance is in line with the current value system, the basis of these actions should not be pursued further. On the other hand, attitudes (regarding for instance issues of equity or the use of authority and power) are often quite clearly visible in students' answers to the situations in the "Interactive examination", which would in fact make it possible to assess them. A way to avoid this dilemma could be to assess students' attitudes, but only for formative purposes. In this way, students who for instance display racist attitudes, or outdated epistemological beliefs in their answers, could be made aware of the fact that their attitudes are not in line with values regarded as appropriate today. This, in turn, would make it possible for the students to reflect on their attitudes; perhaps changing them, but at least to avoid expressing these attitudes or letting them influence their professional decisions.

The next level of inferences that can be made from the scores in the "Interactive examination" is that the scores are generalizable. This means that students' scores should be consistent across assessors, tasks, and occasions (Brennan, 2000; Kane et al., 1999). Such generalizations are supported by the generalizability studies performed. These studies have taken into account some of the major potential threats to generalizability, such as the sampling of tasks and assessors. Still, there is variation due to other sources of error (such as differences in students' writing abilities, differences in "item" difficulty, etc.), which weakens the inferences that can be made from the scores. This is not unique to the "Interactive examination", since such sources of error are present in all assessments (which is a major reason for using different modes of assessment, with presumably different sources of error). Rather, the variance components for students' performance are relatively high, suggesting that the inference is valid, from an observed score to a domain of assessment performances, beyond the specific set of tasks in the examination.

Whether performance extrapolates to the domain of classroom teaching has not been demonstrated, for instance by investigating

whether the students performing well on the "Interactive examination" also perform well as teachers (or vice versa). Instead, this dissertation has so far mainly dealt with the first two levels of inferences, for example by showing that scoring is done via criteria developed in harmony with course objectives, and by indicating that the performance is generalizable across tasks and assessors. The third level of inferences, however, depends to a large extent on how similar the assessment is to the target domain (Kane et al., 1999), and we therefore have to turn to the question of authenticity before discussing this issue further.

Authenticity of the "Interactive examination"

There are two major problems when deciding whether a performance assessment is authentic or not, and these are:

1. The subjectivity of authenticity.
2. That authenticity is a multi-facetted concept.

The notion of subjectivity means that the aspect of authenticity depends on perceptions. As a consequence, while teachers and researcher might categorize a particular assessment as authentic, the students might not. Or while some students might regard the same assessment as being authentic, other students might not. Who, then, is to decide whether an assessment is considered authentic or not? The stance taken in this dissertation is that students' perceptions are most important when validating an assessment for formative purposes. This is based on the assumption that they need to perceive the assessment as authentic in order to apply deep-learning approaches to learning. However, when validating the claim that the "Interactive examination" assesses teacher competency, other arguments besides students' perceptions might be needed. As Gulikers et al. (2004) point out, the concept of authenticity consists of several dimensions that can all vary in their level of authenticity. These dimensions are the task, the physical context, the social context, the form, and the criteria.

In relation to the "Interactive examination", the high level of authenticity is mainly associated with the tasks, which simulate authentic and relevant classroom situations. Still, it could be argued

that the "Interactive examination" is authentic with regard to other dimensions as well:

- The assessment criteria, which are valued in professional settings.
- The assessment form, where students are asked to perform in a way that they are expected to do in real-life settings.
- The social context, which for teachers often means individual work, but also the possibility to consult more experienced colleagues. Lacking in the "Interactive examination", however, is the dynamic interaction with the students.

When considering the physical context, on the other hand, the "Interactive examination" shows little authenticity. Although the time frames and the availability of professional tools provide a certain measure of authenticity, since the "Interactive examination" is performed with the aid of ICT, there is a low level of similarity to professional work space involving face-to-face teaching.

The main reason for not carrying out the examination in the classroom comes from the distinction between which competencies are best assessed in workplace settings, and those more properly assessed in other settings. As discussed in the earlier section on the formulation of criteria for teacher competency, all assessment of professional competency does not have to involve direct participation, and performances other than actual teaching performance can be carried out through more "vicarious means". This means that the classroom is not necessarily the optimal setting for the assessment. In particular, various forms of case-based instruction have been shown to be effective when "transferring" skills from instruction to practice (Michael, Klee, Bransford, & Warren, 1993), and when developing specific skills like self-directed learning (Hmelo-Silver, 2004).

Furthermore, using ICT offers an opportunity to make valid assessments of student competencies in a way that would probably be very time-consuming, labor-intensive, and involve logistical problems, if performed in professional work-space settings instead of through the computer. A paper-and-pencil test, on the other

hand, would provide a much less valid form of assessment (cf. Lam, Williams, & Chua, 2007).

The use of ICT comes with a tradeoff, however, since the physical context of the assessment does not resemble professional work space involving face-to-face teaching, making the assessment less authentic in this regard. Still, restrictions are imposed for different reasons in all assessments, making them narrower than the actual target domain. This means that all performance-assessment tasks are artificial in some ways (Kane et al., 1999), while being relevant in others. Just as noted in relation to the generalizability issues above, this points to the importance of using several different modes of assessment.

A systems approach to assessment

Assessment of competencies is a complex undertaking, and as has been noted, each particular mode of assessment has a number of sources of error and a set of restrictions. This means that any single assessment method is not sufficient to "capture" the full spectrum of skills in the target domain in a credible way. Consequently, a mix of methods must be used, where both traditional tests and newer forms of assessment might be necessary components. For example, in order to assess the competency of a physics teacher, it must be assessed if the student can make use of her subject-matter knowledge when planning and performing instruction. But since it is not possible to assess a broader sample of subject matter knowledge in this way, authentic performance assessments might have to be complemented with more traditional tests. The work by Baartman et al. (2006) supports the notion of such "competency assessment programs", where not all assessments need to fulfill the criteria in their framework to the same extent. This means that some assessments can be authentic with lower levels of reliability, while others are more restricted and have higher levels of reliability. It is important, however, that assessments used for summative purposes are both reliable and valid, since assessments with low levels of validity could be expected to steer student learning towards a more restricted spectrum of learning outcomes. Developing and using assessments that are both reliable and valid is often costly, and an important question is therefore whether all assessments within an

educational program, such as assessments en route, have to be high stakes and summative.

In higher education, due to certification requirements, for instance, assessments tend to be relatively high stakes, demanding high levels of reliability (Mclellan, 2004). But just as assessment validity depends on which interpretations are made from the results, the same is true for reliability. The need for reliability is therefore higher if the decisions made from the assessment cannot be changed, while if decisions can easily be changed, if they turn out to be wrong, there is not the same need to invest in high levels of measurement accuracy (Black, 1998).

This means that, whenever assessments within a program are predominantly formative and embedded in learning activities, assessment resources could be reallocated. This, in turn, could allow for the development of a smaller number of high-quality assessments, which could then be made as reliable as possible (Knight, 2000). Following this line of reasoning, most performance assessments within a teacher-education program should be used for formative purposes, and would therefore not have to meet high demands of reliability. The assessments used for summative purposes, on the other hand, must be both reliable and valid – and, as argued by Wiliam (2008), they should preferably be "synoptic". That assessments are "synoptic" means that they integrate what the students are supposed to have learnt so far in a holistic manner, as opposed to "ticking off" objectives one by one. Synoptic assessments could in this way force both educators and students to adopt a longer-term perspective on learning (see e.g. Brown, 1997; Greatorex & Malacova, 2006).

Assessing teacher competency: Conclusions

The "Interactive examination" is argued to assess teacher competency, since (1) the students act/perform in relation to relevant situations, and are assessed according to valid criteria; (2) the generalization of scores is supported by generalizability studies; and (3) the examination resembles professional settings in several respects. The authenticity of the examination is also confirmed by the students, who perceive the examination as both authentic and meaningful for their future profession. There are, admittedly, limita-

tions to each of these arguments, and no claims are made that the "Interactive examination" should be the only assessment methodology for assessing teacher competency. Instead it is argued that this examination would be useful as one of the assessments in the "competency assessment program" for student teachers as a whole. This in turn means that a battery of different modes of assessment should be used, which in combination address all skills thought to be parts of teacher competency. A central point made here, in this respect, is that all assessments do not need to be summative with high levels of reliability. This is important, not only because achieving high reliability might be costly, but also because high levels of reliability tend to affect validity in a negative way (Brennan, 2000; Dunbar et al., 1991). To avoid such problems, most assessments within a program could instead be formative. On the other hand, assessments used for high-stake decisions need to be reliable, valid, and preferably synoptic, and the "Interactive examination" is argued to satisfy these requirements.

Assessing self-assessment skills

The aim of this dissertation was not only to explore how teacher competency can be assessed in an authentic manner, but also how self-assessment skills can be included in the assessment. On similar grounds as in the previous section, when arguing for the valid assessment of teacher competency, it is possible to question the claim that self-assessment skills are in fact being assessed. Nevertheless, the arguments used in the previous section are only partly applicable to arguments concerning the valid assessment of self-assessment skills. This is because it is not known whether such skills are in fact used by professional teachers, and as a consequence, these skills cannot be argued to be authentic in this regard. There is a tension here, where education can not be limited to mimicking the behavior of existing practice. Rather, education must also strive to renew and potentially improve the profession (see e.g. Svingby, 2003).

Since research on self-assessment does not support the notion of self-assessment skills as a generic ability – rather self-assessment

skills are context dependent and amendable through practice and feedback, just like any other skill – it is important that the assessment of self-assessment skills is implemented in an authentic context, instead of being measured as a context-independent and latent psychological variable. Furthermore, in order to support students' progress, and not only audit it, the self-assessment performed should preferably be qualitative; identifying strengths and weaknesses, as well as indicating how to close the gap between observed and intended performance. These pre-requisites are met in the comparison task in the "Interactive examination".

Still, there is yet another feature present in the "Interactive examination", which separates the assessment of self-assessment skills from the assessment of the other skills aimed for, and that is the professional document.

The professional document

The main purpose of the professional document was to provide an exemplary answer, in relation to which the students could assess their competencies and identify their strengths and shortcomings. In order to do this, however, the students must be able to recognize the professional document as qualitatively different from their own answers – but can they do that?

As was seen in Study II (Jonsson et al., 2007a), even though the task was to identify differences, the majority of the students chose to highlight similarities. Furthermore, some students criticized the professional answer and, in some cases, even regarded their own answers as being qualitatively superior to the professional's. These results could indicate that the students had difficulties in recognizing the differences between their own answers and the professional answer (or possibly that these students tried to surpass the professional in order to get a higher score). On the other hand, those students criticizing the professional were relatively few, and most students, even if emphasizing similarities, managed to identify differences as well. In addition, most differences identified were of a qualitative nature and only rarely referred to quantitative differences, such as the number of words, or to purely stylistic dissimilarities. These results suggest that most of the students were indeed able to recognize differences in quality between their own answers and the

professional answer, albeit not always being able to judge the level of quality. This might, at least partly, be explained by the fact that there are no "correct answers" to the situations, and that the role of the professional document therefore was to analyze the situations as thoroughly, and from as many perspectives, as possible. That the professional documents did not provide a single solution to the problem, but outlined several potential approaches to problem solving, might be interpreted as a sign of ambivalence, rather than a sign of expertise in the eyes of a novice.

Assessing self-assessment skills: Conclusions

In the "Interactive examination", the assessment of self-assessment skills is implemented in an authentic context. Furthermore, the self-assessment performed is of qualitative nature, where the students identify their strengths and weaknesses, as well as formulate their own learning needs. The self-assessment is performed in relation to both criteria and to a professional document, an assessment design which pre-supposes that the students are able to recognize differences in quality between their own and the professional's answer. The results suggest that the students can indeed identify differences in quality, and hence that the assessment of self-assessment skills is valid in this regard.

Supporting student performance

Besides being reliable and valid, assessments should include specific means to help students develop the same competencies that are being aimed for in the assessment. One way of accomplishing this is by making the assessment demands transparent, which in study IV (Jonsson, 2007) was shown to greatly improve student performances. However, in relation to this conclusion the question whether the students are actually *learning* could be raised: Could it be that the students, when supported by a rubric and exemplars, only imitate professional performance, and that the students are in fact performing without learning? For instance, Torrance (2007) and Ecclestone (2007) both argue that transparency in assessment has supported an extensive use of different techniques to improve

achievement, in this way encouraging instrumentalism. According to Torrance, assessment procedures have come to dominate the learning experience, and the notion of "criteria compliance" has replaced the notion of "learning". Wiggins (1998), on the other hand, argues that the very goal of educational assessment is to *educate* and *improve student performance*, and in order to reach this goal, all tasks, criteria, and standards must be transparent to both students and teachers. So how can this controversy be solved?

First, a major problem with investigating learning, is that learning (as opposed to instruction), is unobservable. This means that we can never be sure of when and where a student has actually learned something. Although some authors claim that learning and instruction are like opposite sides of the same coin (see for instance Jank & Meyer, 1997), it is a well known fact that instruction does not always lead to intended learning. Furthermore, intended learning does not always come from instruction (e.g. Kroksmark, 1997). Of course, by using a rigorous experimental design, it could be investigated whether it is *likely* that the students learned specific skills (such as observing and analyzing classroom situations) during the examination or not. But, in fact, it can never be proven that a particular instructional intervention leads to (or causes) the learning intended, and it could consequently be argued that the question of student learning is not possible to answer.

However, since profession-directed education aims for *competency*, it is, when designing formative competency assessments, therefore not necessarily of interest if the students *learn what is intended*[15]. This is because whether they do so or not will in any event have to be evaluated indirectly. What is of interest is if the assessment supports the students in *performing as intended* – where proper performance can be seen as an indicator of learning. For example, results from the "Interactive examination" show that the rubric and exemplars help the students to perform better, suggesting that it is a valid assessment for formative use.

[15] Of course, focusing on performance does not mean that students are not supposed to learn. In contemporary views, learning is not something we can choose not to do, or turn on and off at will. Learning occurs at all times and does not need any reason (Säljö, 2005). Thus students will always learn something, even if it is not necessarily what is instructed (see e.g. Svingby, 1998).

The assessment of adequate performance can not be used exclusively when making inferences about students' competency, however, since the basis of student performance is also an aspect of competency that needs to be assessed (see section *Assessing teacher competency*). By probing students' reasons for acting, it can be judged whether students' performances are merely imitations of professional performance, the results of coincidence or chance, or in fact reflections of "true competency".

A second issue regarding transparency versus learning is that teachers often "teach to the test". This is nothing new. Testing has for some time been argued to give rise to teaching strategies that make test scores go up (Shepard, 2000; Smith & Fey, 2000). This phenomenon (which is also a backwash effect, but at the teacher level), however, is only a problem when the assessment criteria do not reflect central educational outcomes, or the tests are not valid measures of these. If the criteria express, and the tests measure, the outcomes of education that we appreciate as important, while the teachers also teach towards the same end, resulting in the fact that students are able to do the things that we really want them to do – then how could we possibly complain?

So when arguing that formative assessment and transparency improve achievement, but encourage instrumentalism, the underlying problem is in fact neglected, namely that the criteria and achievements are not considered valid. This becomes quite evident when considering the fact that the critique against transparency mentioned above is grounded in research from educational settings which do not have a "coherent curriculum from which assessment is derived; instead it comprises long lists of learning outcomes that are turned into assessment criteria and associated tasks" (Ecclestone, 2007, p. 331).

Supporting student performance: Conclusions

In conclusion, it is argued that the "Interactive examination" is educative in the sense that it supports the students in performing as intended. The students are asked to do the things they are required to do in their future profession (such as analyzing classroom situations and self-assessing their competency), and they are assessed on how well they perform these tasks. Furthermore, the skills of ana-

lyzing classroom situations, and self-assessing in relation to a professional analysis, are not broken down to well defined items, but are kept quite open in an attempt to preserve the complexity and authenticity of these tasks, so that professional performance can be seen as a valid indicator of learning.

Unique features in the "Interactive examination"

The "Interactive examination" is concluded to be valid for both summative and formative purposes, but at the same time constitutes quite a unique methodology, making use of a scoring rubric for teacher competency (including self-assessment skills), for instance, and a specialized ICT solution. This distinctive combination of features has both strengths and weaknesses. A particular strong point is that it provides a powerful environment for assessment, learning, and research, but at the same time it becomes difficult to separate the different parts and investigate their respective contributions to the results. Furthermore, all of these features cannot be expected to be replicated. Bearing this in mind, which seem to bethe "success factors" that might be used to reproduce these results in other settings?

Self-assessment

Something that characterizes effective learners is, according to Boud and Falchikov (1989), that they are aware of their strengths and weaknesses, and also that they can use this knowledge in order to influence their learning in a useful way. In this perspective, self-assessment is not an end in itself, but a means for achieving independence and self-regulation (Wiggins, 1998). Still, self-assessment seems to be an important aspect of effective learning in educational contexts, and Boud and Falchikov argue that successful students have most likely always been good self-assessors. But in order to help less successful students learn how to self-assess, these students need opportunities to practice their self-assessment skills and receive feedback on their performance. The focus on self-assessment is therefore an important feature of the "Interactive examination", and the self-assessment is based on strategies that have been shown

to support students' meta-cognitive learning (i.e. allowing students to practice self-assessment embedded in subject-specific activities in an authentic manner and supporting the self-assessment activities by criteria in the scoring rubric; Boud & Falchikov, op. cit.; Dochy et al., 1999; Jonsson & Svingby, 2007; Topping, 2003).

Although presumably important for student learning, it is not likely that self-assessment has had any significant impact on students' performance in this study. This is because the comparison with examination results can only be performed after the examination (quantitative self-assessment), while the comparison with professional documents can only be performed after the analyses of the classroom situations (qualitative self-assessment). However, by assessing the students in relation to the same criteria at other occasions throughout the teacher-education program, the self-assessment might aid students in developing the necessary skills for achieving independence and self-regulation. The self-assessment scales used in the "Interactive examination" could in this way be used as a kind of "progress indicator" – showing how students progress as professionals, as well as self-assessors. Since there are fewer levels in the rubric as compared to the Likert scales, the qualitative self-assessment might be considered less well suited as a "progress indicator" in this regard. On the other hand, the levels in the rubric are descriptive, which means that the self-assessment criteria could be used as "blueprints" for high quality self-assessment within the community of teaching, providing the students with the necessary "tools of thought" for life-long learning as professional teachers, the main prerequisite being that they compare their own performance to a more professional performance. As the current levels in the rubric are formulated for first-semester students, they could possibly be extended to also include higher levels in the continuum from the novice self-assessor to the expert, in order to use the criteria as "progress indicators" for self-assessment skills.

The scoring rubric

The literature review on rubrics indicated that scoring rubrics potentially have the role of making assessments more reliable and valid, as well as making expectations more transparent to the students, in this way supporting student learning. As discussed pre-

viously, the research on rubrics covered a wide range of contexts and research designs, which means that it is difficult to estimate the generalizability of the findings. Still, the results from the review were used to guide the construction of a rubric in order to optimize it for the intended purposes (i.e. both formative and summative). The assumptions, or guidelines, arrived at were therefore used when constructing the rubric for the "Interactive examination".

For the summative purpose, this means that the rubric was both topic-specific and analytic in order to achieve as high reliability as possible. Also (although not a rubric feature), each student had to do three similar tasks, in order to counterbalance the variability due to tasks. To be able to support valid assessment, the rubric for the "Interactive examination" was developed together with professional teacher educators, and in accordance with the educational objectives of the course in which the examination was implemented. Furthermore, the rubric, together with the examination as a whole, was validated with the aid of a comprehensive framework of validity in Study III (Jonsson et al., 2007a).

For the formative purpose, the rubric was task-related and provided with quality standards. It was distributed to the students well before the examination. Furthermore, both quantitative and qualitative self-assessments for formative purposes are part of the methodology, where the latter is supported by "self-assessment criteria" in the rubric. The methodology thus comprises explicit means to enhance student learning.

Since it is concluded, based on the results from the validation process, that the "Interactive examination" is valid for both its summative as well as its formative purpose, this case indicates that the rubric has been successfully designed in order to aid in fulfilling the intended purposes of the assessment. This does not mean that it has been unambiguously proven that, when combined, these rubric features will always lead to the same results (i.e. validity for both summative and formative purposes), but it provides a thoroughly studied example where it appears that summative and formative purposes can co-exist, as well as a starting point for investigating the assumptions in other contexts.

Transparency

Transparency is argued to be imperative for the formative purpose. If the students do not know what is being assessed, they have to rely on prior experiences of assessment and/or information provided by contextual cues. By making assessment transparent, learning can be more effective, since students can focus their efforts on what is considered important. That this is indeed a successful strategy is evident when reviewing the literature on rubrics and learning (Jonsson & Svingby, 2007), as well as on self-assessment and learning (Dochy et al., 1999; Topping, 2003). Furthermore, Orsmond, Merry, and Reiling (2002) have shown that the use of exemplars can aid the students in improving the quality of their performance. When combining these different means of transparency in the current research, the effectiveness of transparency for improving student performance is further corroborated (Jonsson, 2007).

However, transparency is not only an effective means for improving student performance. It is also a "provider of equality". In studies that are characterized by formative assessment and increased transparency, the difference in attainment between high- and low-performing students is typically reduced (Black & Wiliam, 1998a, 1998b; Sadler & Good, 2006). It was not possible to investigate the difference between high- and low-performing students within the scope of the current research, but the performance of different groups of students was compared, using regression analysis. This analysis showed that there were no significant differences in the results of different groups (e.g. sex, ethnicity, educational status of parents)[16], suggesting an equalizing effect. This effect does not necessarily come from transparency alone, however, but could also be attributed to other factors, such as the use of movies instead of written text. Furthermore, a problem with using regression analysis is that such analysis only shows whether there are differences between the groups. This means that if no differences are found, it has still not been proved that a particular group is not systematically disadvantaged. It could be that, for instance, women

[16] In the 2007 cohort, subject major contributed significantly to the prediction of examination results ($p<.01$), where the Geography majors on the average performed less well than the other students. The difference was relatively small, however, and was not present in the other cohorts.

should be able to perform better, but are somehow disadvantaged, and only perform as well as men. The results from these analyses must therefore be interpreted with caution. Nevertheless, the fact that there was no significant difference between the groups investigated, is in itself a remarkable result, considering the heterogeneity of the sample.

Despite the suggested effectiveness for improving student performance and making assessment more equal, considerable ambiguity subsists in the discussion on transparency. As noted previously, this derives mainly from the fear of pre-defining learning outcomes too strictly and in that way inhibiting creativity and learning. This argument is legitimate, but typically depends on how the transparency is provided. For instance, sharing criteria with the students is not the same thing as telling them *how* to solve the task – only which *qualities* will be assessed (Wiggins, 1998). And, most importantly, the criteria are likely to be used by the teacher when assessing the task (perhaps unconsciously), regardless of whether they are shared with the students or not. It could therefore be argued that the "freedom" provided by not making the criteria explicit is actually a chimera, since what the teacher regards as high- or low-quality work is still pre-defined – explicitly or not.

Besides not specifying the method, another strategy that can be used in order to avoid pre-defining the outcome too narrowly, is to use exemplars representing a wide array of performances (differing both in quality and methods used). In this way the students are confronted with different ways of solving the same task (Wiggins, 1998).

The use of information- and communication technology

As discussed previously, by using ICT there is an opportunity to make valid assessments of student competencies in a way that would not be possible without such technological tools. Furthermore, the same effectiveness, with all students doing the examination at the same time, could probably not be achieved without the aid of computers. This argument is especially important for teacher education, with such a large number of students, and for performance assessments where the students have to perform several tasks in order to achieve acceptable levels of generalizability. In re-

lation to efficiency, another strength of the "Interactive examination" is that the assessment can be operated by one person only, and that all data sources are easily accessible through a database.

Another important aspect of using ICT is the potential for distance education and accessibility. The "Interactive examination" was developed with the explicit aim of being available over the Internet. Internet accessibility was used by a majority of the students who preferred to carry out the examination at home or, in a few cases, from other parts of the world (e.g. Afghanistan, Thailand, and Iceland).

There are also several other ways in which the "Interactive examination" makes use of ICT; for example, by giving access to the professional document directly after the students submit their answers to the personal task, while at the same time saving their answers in a database available to both assessors and researchers.

Unique features: Conclusions

In relation to the summative purpose, it is suggested that the use of a scoring rubric (in combination with letting each student perform three similar tasks) made it possible to achieve higher reliability. Furthermore, the rubric made the assessment more valid, by minimizing the influence of domain-irrelevant factors, and by helping to elicit the performance aimed for.

In relation to the formative purpose, the "Interactive examination" methodology comprises explicit means to enhance student learning, such as a task-related scoring rubric with quality standards, exemplars, and practice in self-assessment. All of these features are aspects of transparency, thought to guide student efforts towards what is considered important. However, while the scoring rubric and the exemplars seem to aid in improving student performance in a more immediate manner, the focus on self-assessment does not necessarily have the same instantaneous effects. Still, self-assessment is seen as an important aspect of student learning, and by self-assessing their competency the students can gain a deeper understanding of the professional qualities aimed for in the examination, as well as learning how to relate them to their own performance as professionals – presumably helping them to better regulate their own learning in the future.

Vital for both the summative and formative purpose is the use of ICT. This is because technology makes it possible to introduce authentic assessment of teacher competency, without necessarily introducing the logistical and practical problems which are often assumed to be associated with such assessments. This authenticity is argued to be essential for assessment validity, as well as for student learning.

Contributions to research

On the basis of the discussion above, it is argued that three main contributions to research have been made. First, by reviewing empirical research on performance assessment and scoring rubrics, a set of assumptions has been reached, concerning how to design authentic assessments that support student learning and simultaneously provide reliable and valid data on student performance. Second, by articulating teacher competency in the form of criteria and standards, as well as displaying it through digital-video simulations, this dissertation has shown that it is possible to assess students' skills in (1) analyzing classroom situations, and in (2) self-assessing their professional competency. Furthermore, it is shown that, by making the assessment demands transparent, student performances are greatly improved. Third, the dissertation provides an illustration of how formative and summative purposes might co-exist within the boundaries of the same (educative) assessment, where teacher competency is assessed in a valid way, without compromising reliability.

Future research

This dissertation acknowledges a number of limitations in the empirical research, and it also raises several new questions, making it important to propose directions for future research. Two major directions for future research are suggested. The first concerns whether student performance in the "Interactive examination" extrapolates to workplace settings, while the second explores how

the implementation of rubrics and transparency affects student motivation and learning as well as teachers' work.

Extrapolation to workplace settings

One of the arguments made in this dissertation is that the students have to perceive the examination as authentic, in order for the assessment to elicit the higher-order thinking required to perform adequately. This, in turn, is assumed to facilitate transfer to the target domain. For instance, it could be hypothesized that the assessment leaves the students with a "tool of thought" for analyzing classroom situations, or for assessing their own competency in relation to more experienced colleagues. Whether this kind of learning actually occurs, and whether the improved performance actually transfers to the target domain, are questions that were not possible to investigate in the current study, but could be addressed through other research designs.

One example of how to investigate whether the skills assessed in the "Interactive examination" transfer to classroom settings could be to use a longitudinal research design, and in this way doing a follow-up study when the students have graduated and start working as teachers. Such a follow-up study could be performed in many different ways. Using for instance standardization as a gradient; at one end of the spectrum simulated situations could be used, just as in the "Interactive examination", but without the rubric. Such a design would make it possible to investigate a large sample of novice teachers, and since the situations would be the same for all participants, this would make data more comparable. At the other end of the standardization spectrum, the study could be performed through observations and interviews, investigating the reasoning used in real-life situations. This design would allow for more in-depth understanding of which "tools of thought" are used, how they are applied in different situations, and how they may have evolved after graduation. Such a design would also bring to light the manners in which novice teachers actually act, where the rationale for their actions can be probed through "stimulated recall" (see e.g. Haglund, 2003) or other similar procedures.

Effects on student motivation and learning, and on teachers' work

The educational potential of transparency could also be further explored. For instance, can the operationalization of transparency made in this dissertation (i.e. rubrics, self-assessment, and exemplars) be implemented in other contexts (such as other higher education institutions or schools), leading to improved performance? Furthermore, how does increased transparency affect students' perceived learning: Do they feel that they learn more, or qualitatively better, or faster, or do they learn totally different things? Another possibility would be to investigate if increased transparency affects the way students experience their work: Do they for instance perceive their work as more meaningful, or more motivating, or perhaps as more strenuous? Teachers' perceptions of their situation is also of great interest, since whether transparency will be implemented on a regular basis would most likely depend on how teachers perceive their work: Does increased transparency lead to increased workload, or only a different kind of workload? Does it make the job easier or more difficult? Does it make it more fun? Does it change the relationship to the students? Does increased transparency affect teachers' view on knowledge?

These questions could be investigated through quantitative means (e.g. pre- and post-tests of student learning) along with combinations of questionnaires and interviews with both students and teachers.

Implications for practice

Implications for practice can be formulated both for the design of performance assessments in general as well as for teacher education.

Implications for the design of performance assessments

Although this dissertation can not offer an unambiguous and conclusive answer as to how performance assessments should be designed in order to support student learning – or how to score such assessments in a credible and trustworthy manner – it does indicate

how rubrics might facilitate such an effort. Rubrics can support student learning by making the assessment transparent, while at the same time making the scoring more reliable and valid. From this proposition, some broadly defined practical implications that might be of use for teachers and/or assessment developers can be presented. These implications correspond to the assumptions that were implemented into the "Interactive examination".

The use of rubrics is likely to increase the reliability of scoring, especially if they are analytic and topic-specific. To reach acceptable levels of reliability, the rubrics should be complemented with benchmarks and/or exemplars and assessor training. Furthermore, rubrics could assume the role of specifying the domain to be assessed, in this way minimizing domain-irrelevant variance. Also, by only rewarding domain-relevant performance, the use of a rubric could help in eliciting the performance required. To be able to support valid assessment in this way, the rubric would have to be developed in accordance with educational objectives or domain theory, and to be validated with the aid of a comprehensive framework of validity. Examples of validity frameworks are Messick's (1996) theory on construct validity or the "Wheel of competency assessment" by Baartman et al. (2006), where the latter includes a wide range of aspects important for formative assessments. In study III (Jonsson et al., 2007a), it is also suggested manners in which the formative purpose can be validated through the use of students' perceptions of the assessment.

In relation to student learning, a rubric could facilitate both feedback and self-assessment. An "instructional rubric" (i.e. used for learning purposes) should be designed according to the principles for formative assessment, such as being task-related and having quality standards. Rubrics should preferably also be combined with some kind of instructional intervention, such as self-assessment, where the students become accustomed to the criteria.

Besides the use of rubrics, the reliability of scoring could be further increased by using two independent assessors and additionally, depending on how the score will be interpreted, the rating scale could be augmented. Perhaps more important, however, is to use more than one task in the assessment, since the variability of stu-

dent performance on performance tasks has been shown to be a major threat to reliability.

Implications for teacher education

A pre-requisite for assessing the professional competency of teachers is the formulation of criteria, since these criteria express what to look for when assessing, and such criteria have been formulated in relation to the "Interactive examination". By developing a rubric, with criteria and standards, this study thus makes a contribution to teacher education by formulating desired outcomes in terms of performance. However, criteria are not only necessary in order to assess whether students have acquired the professional competency aimed for, but also in defining the "competent teacher". In this way, the study also presents a piece of the great "teacher competency jigsaw" to be put under critical scrutiny by researchers and practitioners, and then, hopefully, used in teacher education.

The present study does not only involve the formulation of criteria for assessing teacher competency, but also shows how this competency can be assessed with the aid of ICT. The combination of these features is argued to provide a contribution of high professional significance for how to implement authentic assessment in teacher education, since it has proved possible to incorporate the examination into an existing course within the regular teacher-education program.

Furthermore, criteria formulated for teacher competency are not necessarily limited to assessing individual students. Such criteria can also be used to evaluate the entire education. By assessing the students during their last semester at the teacher-education program, it can be investigated whether the students have indeed profited by the instruction provided and developed the competencies aimed for. Such an evaluation is currently being piloted, where the students who underwent the "Interactive examination" during their first semester in 2003 will take a similar examination during their last semesters in 2007 (primary teachers) and 2008 (secondary teachers).

REFERENCES

Non-English titles have been translated by the author and are marked by square brackets.

Anderson, Jeffrey B. & Freiberg, H. Jerome (1995). Using self-assessment as a reflective tool to enhance the student teaching experience. *Teacher Education Quarterly, 22*, 77-91.

Andrade, Heidi Goodrich (1999a). The role of instructional rubrics and self-assessment in learning to write: A smorgasbord of findings. *Annual meeting of the American Educational Research Association,* Montreal, Canada.

Andrade, Heidi Goodrich (1999b). Student self-assessment: At the intersection of metacognition and authentic assessment. *Annual meeting of the American Educational Research Association,* Montreal, Canada.

Arter, Judith & McTighe, Jay (2001). *Scoring rubrics in the classroom.* Thousand Oaks, CA: Corwin Press.

Baartman, Liesbeth K. J., Bastiaens, Theo J., Kirschner, Paul A., & Van der Vleuten, Cees P. M. (2006). The wheel of competency assessment: Presenting quality criteria for Competency Assessment Programs. *Studies in Educational Evaluation, 33*, 153-177.

Baartman, Liesbeth K. J., Bastiaens, Theo J., Kirschner, Paul A., & Van der Vleuten, Cees P. M. (2007). Teachers' opinions on quality criteria for Competency Assessment Programs. *Teaching and Teacher Education, 23,* 857-867.

Baartman, Liesbeth K. J., Bastiaens, Theo J., Kirschner, Paul A., & Van der Vleuten, Cees P. M. (2008). Evaluating assessment quality in competence-based education: A qualitative comparison of two frameworks, *Educational Research Review, 2*, 114-129.

Baker, Eva L., Abedi, Jamal, Linn, Robert L., & Niemi, David (1995). Dimensionality and generalizability of domain-independent performance assessments. *Journal of Educational Research, 89*, 197-205.

Baxter, Gail P. & Glaser, Robert (1998). Investigating the cognitive complexity of science assessments. *Educational Measurement: Issues and Practice, 17*, 37-45.

Berliner, David C. (1986). In pursuit of the expert pedagogue. *Educational Researcher, 15*, 5-13.

Berliner, David C. (2004). Describing the behavior and documenting the accomplishments of expert teachers. *Bulletin of Science, Technology & Society, 24*, 200-212.

Biggs, John (1987). *Student approaches to learning.* Melbourne: Australian Council for Educational Research.

Biggs, John (1996). Enhancing teaching through constructive alignment. *Higher Education, 32*, 347-364.

Biggs, John (1999). *Teaching for Quality Learning at University*. Buckingham: SRHE and Open University Press.

Birenbaum, Menucha (2003). New insights into learning and teaching and their implications for assessment. In Mien Segers, Filip Dochy, & Eduardo Cascallar (Eds.), *Optimizing new modes of assessment: In search of qualities and standards* (pp. 13-36). Dordrecht: Kluwer Academic Publishers.

Birenbaum, Menucha, Breuer, Klaus, Cascallar, Eduardo, Dochy, Filip, Dori, Yehudit, Ridgway, Jim, Wiesemes, Rolf, & Nickmans, Goele (2006). A learning integrated assessment system. *Educational Research Review, 1*, 61-67.

Black, Paul (1998). *Testing: Friend or Foe?* London: Falmer press.

Black, Paul, Harrison, Christine, Lee, Clare, Marshall, Bethan, & Wiliam, Dylan (2003). *Assessment for learning. Putting it into practice.* Berkshire: Open University Press.

Black, Paul & Wiliam, Dylan (1998a). Assessment and classroom learning. *Assessment in Education, 5*, 7-74.

Black, Paul & Wiliam, Dylan (1998b). Inside the black box: Raising standards through classroom assessment. *Phi Delta Kappan, 80*, 139-148.

Borsboom, Denny, Mellenbergh, Gideon J., & van Heerden, Jaap (2004). The concept of validity. *Psychological Review, 111*, 1061-1071.

Boud, David & Falchikov, Nancy (1989). Quantitative studies of self-assessment in higher education: a critical analysis of findings. *Higher Education, 18*, 529-549.

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

Brown, Alan (1997). Possible role for synoptic assessment within vocational education pathways. In Sabine Manning (Ed.), *Qualifications with a Dual Orientation towards Employment and Higher Education. A Collaborative Investigation of Selected Issues in Seven European Countries. INTEQUAL*

Report II (pp. 49-80). Berlin: WIFO (Research Forum Education and Society).

Brown, Gavin T. L., Glasswell, Kath, & Harland, Don (2004). Accuracy in the scoring of writing: Studies of reliability and validity using a New Zealand writing assessment system. *Assessing Writing, 9*, 105-121.

Brown, George, Bull, Joanna, & Pendlebury, Malcolm (1997). *Assessing student learning in higher education.* London: Routledge.

Busching, Beverly (1998). Grading inquiry projects. *New directions for teaching and learning*, 89-96.

Carter, Kathy, Cushing, Katherine, Sabers, Donna, Stein, Pamela, & Berliner, David (1988). Expert-novice differences in perceiving and processing visual classroom information. *Journal of Teacher Education, 39*, 25-31.

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

Denner, Peter R., Salzman, Stephanie A., & Harris, Larry B. (2002). Teacher work sample assessment: An accountability method that moves beyond teacher testing to the impact of teacher performance on student learning, *Annual meeting of the American Association of Colleges for Teacher Education*, New York, NY, USA.

Dewey, John (1980). *Art as experience*. New York: Perigee books.

Dilthey, Wilhelm (1996). *Hermeneutics and the study of history.* Princeton, NJ: Princeton University Press.

Dochy, Filip, Segers, Mien, & Sluijsmans, Dominique (1999). The use of self-, peer and co-assessment in higher education: a review. *Studies in Higher Education, 24*, 331-350.

Dreyfus, Hubert L. & Dreyfus, Stuart E. (1988). *Mind over machine.* New York: Free Press.

Duke, Bryan L. (2003). *The influence of using cognitive strategy instruction through writing rubrics on high school students' writing self-efficacy, achievement goal orientation, perceptions of classroom goal structures, self-regulation, and writing achievement.* Unpublished doctoral dissertation, University of Oklahoma.

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

Dysthe, Olga, Engelsen, Knut Steinar, Madsen, Tjalve, & Wittek, Line (2008). A theory-based discussion of assessment criteria. The balance between explicitness and negotiation. In Anton Havnes & Liz McDowell (Eds.), *Balancing Dilemmas in Assessment and Learning in Contemporary Education* (pp. 121-131). New York, NY: Routledge.

Ecclestone, Kathryn (2007). Commitment, compliance and comfort zones: the effects of formative assessment on vocational education students' learning careers. *Assessment in Education: Principles, Policy & Practice, 14*, 315-333.

Eisner, Elliot (1991). Taking a Second Look: Educational Connoisseurship Revisited. In Milbrey W. McLaughlin & D. C. Phillips (Eds.), *Yearbook of the National Society for the Study of Education. Evaluation and education at quarter century* (pp. 169-187). Chicago, IL: The National Society for the Study of Education.

Elliott, John (1991). A model of professionalism and its implications for teacher education. *British Educational Research Journal, 17*, 309-318.

Entwistle, Noel J. (1991). Approaches to learning and perceptions of the learning environment. Introduction to special issue. *Higher Education, 22*, 201-204.

Entwistle, Noel J. & Ramsden, Paul (1983). *Understanding Student Learning.* London & Canberra: Croom Helm.

Entwistle, Noel J. & Peterson, Elizabeth R. (2004). Conceptions of learning and knowledge in higher education: Relationships with study behaviour and influences of learning environments. *International Journal of Educational Research, 41*, 407–428.

Evertson, C., Hawley, W., & Zlotnik, M. (1985). Making a difference in educational quality through teacher education. *Journal of Teacher Education, 36*, 2-12.

Falchikov, Nancy & Boud, David (1989). Student self-assessment in higher education: A meta-analysis. *Review of Educational Research, 59*, 395-430.

Flavell, John H., Miller, Patricia H., & Miller, Scott A. (1993). *Cognitive development.* Englewood Cliffs, NJ: Prentice-Hall.

Flyvbjerg, Bent (2001). *Making social science matter: Why social inquiry fails and how it can succeed again.* Cambridge: Cambridge University Press.

Folkesson, Anne-Marie & Svingby, Gunilla (Unpublished manuscript). *Teacher students' self-report constructs. Validation and relations to demographics.*

Fransson, A. (1977). On qualitative differences in learning: IV – Effects of intrinsic motivation and extrinsic test anxiety on process and outcome. *British Journal of Educational Psychology, 47*, 244-257.

Frederiksen, John R. & Collins, Allan (1989). A systems approach to educational testing. *Educational Researcher, 18*, 27-32.

Gao, Xiaohong, Shavelson, Richard J., & Baxter, Gail P. (1994). Generalizability of large-scale performance assessments in science: Promises and problems. *Applied Measurement in Education, 7*, 323-342.

Gearhart, Maryl, Herman, Joan L., Novak, John R., & Wolf, Shelby A. (1995). Toward the instructional utility of large-scale writing assessment: Validation of a new narrative rubric. *Assessing Writing, 2*, 207-242.

Gielen, Sarah, Dochy, Filip, & Dierick, Sabine (2003). Evaluating the consequential validity of new modes of assessment: The influence of assessment on learning, including pre-, post-, and true assessment effects. In Mien Segers, Filip Dochy, & Eduardo Cascallar (Eds.), *Optimizing new modes of assessment: In search of qualities and standards* (pp. 37-54). Dordrecht: Kluwer Academic Publishers.

Giertz, Birgitta (2003). Att bedöma pedagogisk skicklighet – går det? [Assessing teaching proficiency – can it be done?]. Rapportserie från Avdelningen för utveckling av pedagogik och interaktivt lärande. Rapport nr 2. Uppsala: Uppsala universitet.

Gijbels, David & Dochy, Filip (2006). Students' assessment preferences and approaches to learning: can formative assessment make a difference? *Educational Studies, 32*, 399-409.

Gijbels, David, Van de Watering, Gerard, Dochy, Filip, & Van den Bossche, Piet (2005). The relationship between students' approaches to learning and the assessment of learning outcomes. *European Journal of Psychology of Education, 20*, 327-341.

Gipps, Caroline (2001). Sociocultural aspects of assessment. In Gunilla Svingby & Sofia Svingby (Eds.), *Bedömning av kunskap och kompetens* [Assessment of knowledge and competence] (pp. 15-67). Stockholm: Lärarhögskolan i Stockholm, PRIM-gruppen.

Gonczi, Andrew (1994). Competency based assessment in the professions in Australia. *Assessment in Education: Principles, Policy & Practice, 1*, 27-41.

Greatorex, Jackie & Malacova, Eva (2006). Can different teaching strategies or methods of preparing pupils lead to greater improvements from GCSE to A level performance? *Research Papers in Education, 21*, 255–294.

Gulikers, Judith T. M., Bastiaens, Theo J., & Kirschner, Paul A. (2004). A five-dimensional framework for authentic assessment. *Educational Technology Research & Development, 52*, 67-86.

Haglund, Björn (2003). Stimulated recall. Några anteckningar om en metod att generera data [Stimulated recall. Some notes on a method to generate data]. *Pedagogisk Forskning i Sverige, 8*, 145–157.

Hamilton, Laura S., Nussbaum, E. Michael, & Snow, Richard E. (1997). Interview procedures for validating science assessments. *Applied Measurement in Education, 10*, 181-200.

Hammerness, Karen, Darling-Hammond, Linda, Bransford, John, Berliner, David, Cochran-Smith, Marilyn, McDonald, Morva, & Zeichner, Kenneth (2005). How teachers learn and develop. In Linda Darling-Hammond &

John Bransford (Eds.), *Preparing teachers for a changing world* (pp. 358-389). San Francisco, CA: Jossey-Bass.

Havnes, Anton (2008). Assessment. A boundary object linking professional education and work? In Anton Havnes & Liz McDowell (Eds.), *Balancing Dilemmas in Assessment and Learning in Contemporary Education* (pp. 101-114). New York, NY: Routledge.

Hmelo-Silver, Cindy E. (2004). Problem-based learning: What and how do students learn? *Educational Psychology Review, 16*, 235-266.

Huxham, Mark (2007). Fast and effective feedback: are model answers the answer? *Assessment & Evaluation in Higher Education, 32*, 601-611.

Jank, Werner & Meyer, Hilbert (1997). Nyttan av kunskaper i didaktisk teori [The usefulness of knowledge in didactic theory]. In Michael Uljens (Ed.), *Didaktik* (pp. 17-34). Lund: Studentlitteratur.

Jonsson, Anders (2007). The use of transparency in the "Interactive examination" for student teachers. *AEA Europe Conference,* Stockholm, Sweden.

Jonsson, Anders, Baartman, Liesbeth K. J., & Lennung Sven A. (2007a). Estimating the quality of performance assessments: The case of an "Interactive examination" for teacher competency. *EARLI Conference,* Budapest, Hungary.

Jonsson, Anders, Mattheos, Nikos, Svingby, Gunilla, & Attström, Rolf (2007b). Dynamic assessment and the "Interactive examination". *Educational Technology & Society, 10,* 17-27.

Jonsson, Anders & Svingby, Gunilla (2007). The use of scoring rubrics: Reliability, validity and educational consequences. *Educational Research Review, 2*, 130-144.

Kane, Michael, Crooks, Terence, & Cohen, Allan (1999). Validating measures of performance. *Educational Measurement: Issues and Practice, 18*, 5-17.

Kluger, Avraham N. & DeNisi, Angelo (1996). The effects of feedback interventions on performance: A historical review, a meta-analysis, and a preliminary feedback intervention theory. *Psychological Bulletin, 119*, 254-284.

Knight, Peter T. (2000). The value of a program-wide approach to assessment. *Assessment & Evaluation in Higher Education, 25*, 237-251.

Knight, Peter, Tait, Jo, & Yorke, Mantz (2006). The professional learning of teachers in higher education. *Studies in Higher Education, 31*, 319–339.

Kroksmark, Tomas (1997). Undervisningsmetodik som forskningsområde [Didactics as an area of research]. In Michael Uljens (Ed.), *Didaktik* (pp. 77-97). Lund: Studentlitteratur.

Lam, Wing, Williams, Jeremy B., & Chua, Alton Y. K. (2007). E-xams: harnessing the power of ICTs to enhance authenticity. *Educational Technology & Society, 10*, 209-221.

Lave, Jean & Wenger, Etienne (1991). *Situated learning: legitimate peripheral participation*. Cambridge: Cambridge University Press.

Lin, Sunny S. J. (1999). Looking for the prototype of teaching expertise: An initial attempt in Taiwan. *Annual Meeting of the American Educational Research Association*, Boston, MA, USA.

Lindblom-Ylänne, Sari (2003). Broadening an understanding of the phenomenon of dissonance. *Studies in Higher Education, 28*, 63-77.

Lindström, Lars (2001). *Från novis till mästare. En studie av bedömningskriterier i slöjd* [From novice to expert. A study of assessment criteria in handycraft]. Stockholm: Lärarhögskolan i Stockholm.

Lindström, Lars (2006). Creativity: What Is It? Can You Assess It? Can It Be Taught? *Journal of Art & Design Education, 25*, 53-66.

Lindström, Lars (2008). Assessing craft and design. Conceptions of expertise in education and work. In Anton Havnes & Liz McDowell (Eds.), *Balancing Dilemmas in Assessment and Learning in Contemporary Education* (pp. 61-72). New York, NY: Routledge.

Lindström, Lars, Ulriksson, Leif, & Elsner, Catharina (1999). *Portföljvärdering av elevers skapande i bild* [Portfolio assessment of pupils' creative work in art]. Stockholm: Skolverket.

Linn, Robert L., Baker, Eva. L., & Dunbar, Stephen B. (1991). Complex, performance-based assessment: Expectations and validation criteria. *Educational Researcher, 20*, 15-21.

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

Lonka, Kirsti, Olkinuora, Erkki, & Mäkinen Jarkko (2004). Aspects and prospects of measuring studying and learning in higher education. *Educational Psychology Review, 16*, 301-323.

Lundeberg, Mary A. & Fox, Paul W. (1991) Do laboratory findings on test expectancy generalize to classroom outcomes? *Review of Educational Research, 61*, 94-106.

Malmberg, Claes (2006). *Kunskapsbygge på nätet. En studie av studenter i dialog* [Net-based knowledge building. A study of students in dialogue.]. Unpublished doctoral dissertation, Malmö University, Sweden.

Malmberg, Claes & Svingby, Gunilla (2004). Students' dialogues as contributions in education for sustainable development. In Per Wickenberg, Harriet Axelsson, Lena Fritzén, Gustav Helldén, & Johan Öhman (Eds.), *Learning to change our world* (pp. 263-274). Lund: Studentlitteratur.

Marton, Ference & Säljö, Roger (1984). Approaches to learning. In Ference Marton, Dai Hounsell, & Noel Entwistle (Eds.), *The experience of learning* (pp. 36-55). Edinburgh: Scottish Academic Press.

Mattheos, Nikos, Nattestad, Anders, Christersson, Cecilia, Jansson, Henrik, & Attström, Rolf (2004a). The effects of an interactive software application on the self-assessment ability of dental students. *European Journal of Dental Education, 8*, 97-104.

Mattheos, Nikos, Nattestad, Anders, Falk Nilsson, Eva, & Attström, Rolf (2004b). The interactive examination: Assessing students' self-assessment ability. *Medical Education, 38*, 378-389.

Mclellan, Effie (2004). How convincing is alternative assessment for use in higher education? *Assessment & Evaluation in Higher Education, 29,* 311-321.

McMartin, Flora, McKenna, Ann, & Youssefi, Ken (2000). Scenario assignments as assessment tools for undergraduate engineering education. *IEEE Transactions on Education, 43*, 111-120.

McMillan, James H. (2004). *Educational research: Fundamentals for the consumer*. Boston, MA: Pearson Education.

Messick, Samuel (1996). Validity of performance assessments. In Gary W. Phillips (Ed.), *Technical issues in large-scale performance assessment* (pp. 1-18). Washington, DC: National Center for Education Statistics.

Messick, Samuel (1998). Test validity: A matter of consequence. *Social Indicators Research, 45*, 35–44.

Metcalf, Kim K., Ronen Hammer, M. A., & Kahlich, Pamela A. (1996). Alternatives to field-based experiences: The comparative effects of on-campus laboratories. *Teaching & Teacher Education, 12*, 271-283.

Michael, Ann L., Klee, Thomas, Bransford, John D., & Warren, Steven F. (1993). The transition from theory to therapy: Test of two instructional methods. *Applied Cognitive Psychology, 7*, 139-154.

Miller, David M. (1998). *Generalizability of performance-based assessments.* Washington, DC: Council of Chief State School Officers.

Miller, David M. (1999). *Teacher uses and perceptions of the impact of statewide performance-based assessments.* Washington, DC: Council of Chief State School Officers.

Miller, David M. & Linn, Robert L. (2000). Validation of performance-based assessments. *Applied Psychological Measurement, 24*, 367-378.

Moskal, Barbara M., & Leydens, Jon A. (2000). Scoring rubric development: Validity and reliability. *Practical Assessment, Research & Evaluation, 7*, 71-81.

Mullen, Yvonne K. (2003). *Student improvement in middle school science.* Unpublished master dissertation, University of Wisconsin.

Orsmond, Paul & Merry, Stephen (1996). The importance of marking criteria in the use of peer assessment. *Assessment & Evaluation in Higher Education, 21*, 239-250.

Orsmond, Paul, Merry, Stephen, & Reiling, Kevin (1997). A study in self-assessment: Tutor and students' perceptions of performance criteria. *Assessment & Evaluation in Higher Education, 22*, 357-370.

Orsmond, Paul, Merry, Stephen, & Reiling, Kevin (2002). The use of exemplars and formative feedback when using student derived marking criteria in peer and self-assessment. *Assessment & Evaluation in Higher Education*, 27, 309-323.

Osana, Helena P. & Seymour, Jennifer R. (2004). Critical thinking in preservice teachers: A rubric for evaluating argumentation and statistical reasoning. *Educational Research and Evaluation, 10*, 473-498.

Perlman, Carole C. (2003). Performance assessment: Designing appropriate performance tasks and scoring rubrics. In Janet E. Wall & Garry R. Walz (Eds.), *Measuring Up: Assessment Issues for Teachers, Counselors, and Administrators.* Greensboro, NC: ERIC Counseling and Student Services Clearinghouse.

Piscitello, Mary E. (2001). *Using rubrics for assessment and evaluation in art.* Unpublished master dissertation. Saint Xavier University: Chicago, IL.

Polanyi, Michael (1983). *The tacit dimension*. Gloucester: Peter Smith.

Popham, James W. (1997). Consequential validity: Right concern – wrong concept. *Educational Measurement: Issues and Practice, 16*, 9-13.

Ramos, Kathleen D., Schafer, Sean, & Tracz, Susan M. (2003). Validation of the Fresno test of competence in evidence based medicine. *British Medical Journal, 326*, 319-321.

Ruiz-Primo, Maria A., Li, Min, Ayala, Carlos, & Shavelson, Richard J. (2004). Evaluating students' science notebooks as an assessment tool. *International Journal of Science Education, 26*, 1477-1506.

Sadler, Philip M. & Good, Eddie (2006). The impact of self- and peer-grading on student learning. *Educational assessment, 11*, 1-31.

Sadler, Royce D. (1987). Specifying and promulgating achievement standards. *Oxford Review of Education, 13*, 191-209.

Sadler, Royce D. (1989). Formative assessment and the design of instructional systems. *Instructional Science, 18*, 119-144.

Sadler, Royce D. (1998). Formative assessment: Revisiting the territory. *Assessment in Education: Principles, Policy & Practice, 5*, 77-84.

Schacter, John & Thum, Yeow Meng (2004). Paying for high- and low-quality teaching. *Economics of Education Review 23*, 411–430.

Schamber, Jon F. & Mahoney, Sandra L. (2006). Assessing and improving the quality of group critical thinking exhibited in the final projects of collaborative learning groups. *Journal of General Education, 55*, 103-137.

Schön, Donald A. (1983). *The reflective practitioner: How professionals think in action.* Aldershot: Ashgate.

Scouller, Karen (1998). The influence of assessment method on students' learning approaches: Multiple choice question examination versus assignment essay. *Higher Education, 35*, 453-472.

Segers, Mien, Dochy, Filip, & Cascallar, Eduardo (2003). The era of assessment engineering: Changing perspectives on teaching and learning and the role of new modes of assessment. In Mien Segers, Filip Dochy, & Eduardo Cascallar (Eds.), *Optimizing new modes of assessment: In search of qualities and standards* (pp. 1-12). Dordrecht: Kluwer Academic Publishers.

Segers, Mien, Nijhuis Jan, & Gijselaers, Wim (2006). Redesigning a learning and assessment environment: The influence on students' perceptions of assessment demands and their learning strategies. *Studies in Educational Evaluation, 32*, 223-242.

SFS 2001:23. *Förordning om ändring i högskoleförordningen (1993:100)* [Statute on change in the statutory regulations of higher education institutions]. Appendix 2[4].

Shavelson, Richard J. & Webb, Noreen M. (1991). *Generalizability theory: A primer.* Newbury Park, CA: Sage.

Shepard, Lorrie A. (2000). The role of assessment in a learning culture. *Educational Researcher, 29*, 4-14.

Shepard, Lorrie A. (2002). The role of classroom assessment in teaching and learning. In Virginia Richardson (Ed.), *Handbook of Research on Teaching, 4th ed.* (pp. 1066-1101). Washington, DC: American Educational Research Association.

Smith, Mary Lee & Fey, Patricia (2000). Validity and accountability of high-stakes testing. *Journal of Teacher Education, 51*, 334-344.

Spurling, Steven D. (1979). Contextualized and decontextualized tests. *TESOL Quarterly, 13*, 597-604.

Stake, Robert E. (1994). Case studies. In Norman K. Denzin & Yvonna S. Lincoln (Eds.), *Handbook of qualitative research* (pp. 236-247). Thousand Oaks, CA: Sage.

Stemler, Steven E. (2004). A comparison of consensus, consistency, and measurement approaches to estimating interrater reliability. *Practical Assessment, Research & Evaluation, 9*.

Struyven, Katrien, Dochy, Filip & Janssens, Steven (2005). Students' perceptions about new modes of assessment in higher education: a review. *Assessment & Evaluation in Higher Education, 30*, 325–341.

Svingby, Gunilla (1998). *Proven, kunskapen och undervisningen. Samhällsorienterande ämnen. Rapport nr. 138.* [Tests, Teaching and Knowledge in Social Studies. The National Evaluation of Social Studies]. Stockholm: Skolverket.

Svingby, Gunilla (2003). *Accessibility and Learning in Higher Education: Learning and New Media. Final report*. Malmö: Malmö University.

Svingby, Gunilla (2005). Kunskaps- och verktygsbanken – Ett digitalt verktyg för att utveckla lärare och lärarstudenters professionella kompetens [The knowledge and competency bank. A digital tool to help teachers and teacher students develop professional competency]. *Application submitted to the KK-foundation.* Malmö: Malmö University.

Säljö, Roger (1975). *Qualitative differences in learning as a function of the learner's conception of the task*. Unpublished doctoral dissertation. Gothenburg: University of Gothenburg.

Säljö, Roger (2005). *Lärande och kulturella redskap: om lärprocesser och det kollektiva minnet* [Learning and cultural tools: About learning processes and the collective memory]. Stockholm: Norstedts akademiska förlag.

Tabachnik, B., Popkewitz, T., & Zeichner, K. (1979-80). Teacher education and the professional perspectives of student teachers. *Interchange, 10*, 12-29.

Taconis, R., Van der Plas, P. & Van der Sanden, J. (2004). The development of professional competencies by educational assistants in school-based teacher education, *European Journal of Teacher Education, 27(2)*, 215-240.

Topping, Keith (2003). Self and peer assessment in school and university: Reliability, validity and utility. In Mien Segers, Filip Dochy, & Eduardo Cascallar (Eds.), *Optimizing new modes of assessment: In search of qualities and standards* (pp. 55-87). Dordrecht: Kluwer Academic Publishers.

Torrance, Harry (2007). Assessment as learning? How the use of explicit learning objectives, assessment criteria and feedback in post-secondary education and training can come to dominate learning. *Assessment in Education: Principles, Policy & Practice, 14*, 281-294

Toth, Eva Erdosne, Suthers, Daniel D., & Lesgold, Alan M. (2002). "Mapping to know": The effects of representational guidance and reflective assessment on scientific inquiry. *Science Education, 86*, 264-286.

Tunstall, Pat & Gipps, Caroline (1996). Teacher feedback to young children in formative assessment: A typology. *British Educational Research Journal, 22*, 389-405.

van den Berg, Ellen (2001). An exploration of the use of multimedia cases as a reflective tool in teacher education. *Research in Science Education, 31*, 245–265.

Wells, Gordon (2001). The case for dialogic inquiry. In Gordon Wells (Ed.), *Action, Talk, & Text: Learning and teaching through inquiry*. New York, NY: Teachers College Press.

Wenger, Etienne (1998). *Communities of practice: learning, meaning, and identity*. Cambridge: Cambridge University Press.

Wertsch, James V. (1991). *Voices of the mind: a sociocultural approach to mediated action.* London: Harvester Wheatsheaf.

Wertsch, James V. (1998). *Mind as action.* New York, Oxford: Oxford University Press.

Wiiand, Towe (1998). *Examinationen i fokus. Högskolestudenters lärande och examination – en litteraturöversikt* [Focusing on the examination. University students' learning and examination – a literature review]. Rapportserie från Enheten för utveckling och utvärdering. Rapport nr 14. Uppsala: Uppsala universitet.

Wiiand, Towe (2005). *Examinationen som vägvisare. Högskolestudenters upplevelse av examination i ett longitudinellt perspektiv* [The examination as a guide. University students' perceptions of examinations in a longitudinal perspective]. Rapportserie från Avdelningen för utveckling av pedagogik och interaktivt lärande. Rapport nr 5. Uppsala: Uppsala universitet.

Wiggins, Grant (1998). *Educative assessment.* San Francisco, CA: Jossey-Bass.

Wiliam, Dylan (2008). Balancing dilemmas. Traditional theories and new applications. In Anton Havnes & Liz McDowell (Eds.), *Balancing Dilemmas in Assessment and Learning in Contemporary Education* (pp. 267-281). New York, NY: Routledge.

Williams, Lori & Rink, Judith (2003). Teacher competency using observational scoring rubrics. *Journal of Teaching in Physical Education, 22*, 552-572.

Yeh, Yu-Chu. (2004). Nurturing reflective teaching during critical-thinking instruction in a computer simulation program. *Computers & Education, 42*, 181-194.

Yerrick, Randy, Ross, Donna, & Molebash, Philip (2005). Too close for comfort: Real-time science teaching reflections via digital video editing. *Journal of Science Teacher Education, 16*, 351–375.

Zary, Nabil, Johnson, Gunilla, Boberg, Jonas, & Fors, Uno (2006). Development, implementation and pilot evaluation of a Web-based Virtual Patient Case Simulation environment – Web-SP. *BMC Medical Education, 6.*

APPENDICES

Appendix A. The "Interactive examination"

The examination starts with a *quantitative self-assessment*, where the students assess their competency on a scale from 1 to 6. The self-assessment questions are equivalents to the criteria in the scoring rubric (see Table A1).

Table A1. Example of a self-assessment question and a criterion in the rubric.

Self-assessment question	*Criterion in rubric*
How do you judge your competency in describing classroom situations without prejudice? (1 = poor, 6 = excellent)	The description is not prejudiced.

In the next step, the *personal task*, the students watch three different short movie sequences. For each movie, they have to answer three global questions (Observation, Analysis, and Taking action), which are sequenced by the software. Figure A1 shows a screenshot of a movie sequence, and Figure A2 gives an example of a student answer to this particular movie sequence.

After submitting their answers to one of the personal tasks, the students can access a professional answer to the same movie sequence. Figure A3 shows what the student interface looks like and how the students can access the professional document.

Figure A4 gives an example of a student *comparison with a professional* answer.

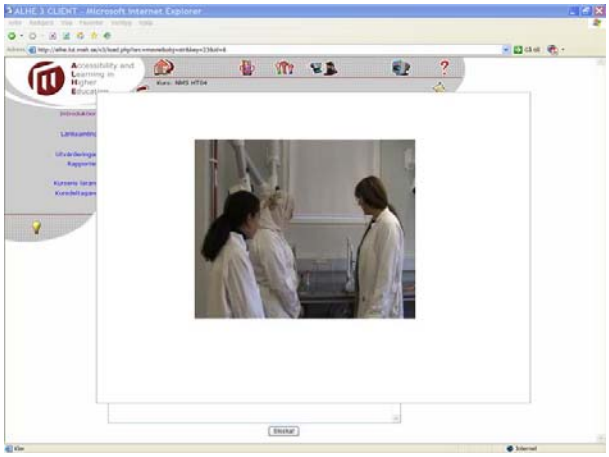


Figure A1. Screenshot from the student interface (in the older version) showing how one of the movies is played in the browser. The situation displays how a Muslim girl, for safety reasons, is requested to remove her veil in a Science class.

The chemistry teacher explains to the students what they are supposed to do during class. The teacher tells them that they are going to be working with an open flame, and for safety reasons she wants the Muslim girl to remove her veil, so it will not catch fire. But since the girl has to wear a veil according to her religion, she cannot take it off. In today's multicultural schools, the teacher should take students' religious beliefs into consideration. Instead of forcing the girl to take her veil off, the teacher should aid in dividing the work between the girls so both of them can participate in the experiment, and in this way take into consideration that the girl cannot perform all parts in this particular experiment. It is not acceptable for the teacher to threaten the girl by saying that her grade might be endangered if not taking the veil off. The teacher shows, through her actions, a "we-versus-them" behavior. By "we" is meant what is supposed to be our Swedish standards and values, and by "them" are those who do not belong (Trondman & Bunar, 2001). The teacher also shows hostility towards foreigners and is unsympathetic to the culture and the religious beliefs of other individuals. The consequences might be that the student perceives the situation as depreciating and alienating. As a teacher it is important to get all students to feel that they are active participants in school (lecture on "the process of reading and writing" by A.A.).

Figure A2. An example of a "typical" (i.e. neither representing very low- nor very high performance) student answer to the movie showed in Figure A1. The answer is from the Analysis question.

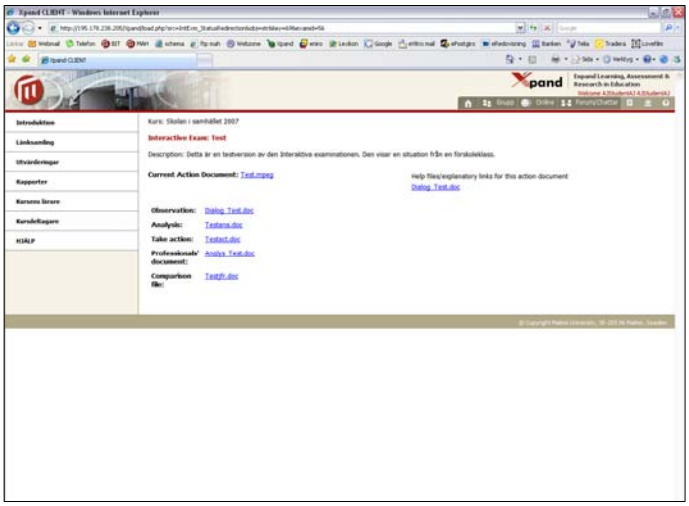


Figure A3. Screenshot from the student interface (in the newer version) showing how the student can access different resources (such as the movie sequence, dialogue in text format, own submitted answers, and professional document) through hyperlinks.

Observation

The difference between my observation and the professional's is mainly due to language, i.e. which words and expressions were used. One thing that the professional has written about, which I have not, is that there are no other students visible in the sequence, and it is not apparent whether other students can hear the conversation. I took for granted that no other students were present, since I only described who were present in the sequence. It did not occur to me that other students could hear the conversation.

Analysis

Both the professional and I have written about safety, which according to the teacher in the movie is the reason why the student should remove the veil. The professional writes about the student arguing against the teacher, which was something that I did not see in the situation. Furthermore, the teacher's threat regarding the student's grade was mentioned by both the professional and me, even though the professional has written about it more extensively than I have, and with references to course objectives, something that I do not have enough knowledge about. I perceived the teacher's threat about grades as a way to show her negative attitude towards this religious artifact in general. I thought the discussion about grades was based on communicational problems between student and teacher, something that the professional has not written about.

Regarding the symbolism of the veil, the professional has done a more thorough analysis as to how the teacher's attitude affects the student. This was something that I wanted to write about, but did not feel that I could on the basis of the course literature. Both the professional and I agreed on that the teacher should start thinking about her view on the student and her choice of wearing a veil.

As to the student's and the teacher's behavior, I do not feel that I have the proper knowledge base to analyze this in an adequate manner.

Figure A4a. An example of a student comparison with the professional document (continued on next page).

Taking action

The professional thinks that the teacher does not have the readiness to handle the situation, which I agree with. Starting from safety regulations, we both think it would be possible to get fire-proof covers for both hair and veil. The professional has suggested actions in a longer perspective, pointing out that the school should look over the regulations and adapt them to a multicultural society.

I do not see the other student performing the experiment as a solution to the problem, but I understand that it might be necessary in this particular situation. Both the professional and I recommend a conversation outside the classroom in order to take care of the problem.

The professional has written about problems with authority, which I have not done at any length. I thought the teacher used her authority in a negative way, but I have not written any more than that. The professional has suggested concrete solutions which I think are good, but I do not know enough to recommend such solutions myself. We both write that the teacher should start thinking about her view on peoples' religious and cultural values. A teacher must not discriminate against any students and should have advance planning for similar conflicts in the future.

Shortcomings in my competency as a teacher and my learning needs:

I took it for granted that no other students were present – which I see as a shortcoming. A teacher should not take things for granted, but analyze each situation thoroughly.

I did not think about whether other students heard the conversation in the sequence. – I want to be able to see the bigger picture, and as mentioned above, do not want to take things for granted. Even if the movie shows one thing, this does not necessarily have to be the only truth.

Figure A4b. An example of a student comparison with a professional answer (continued). Example continues on next page.

The student arguing against the teacher, something that I did not see in the situation. – Difficult to know why I did not see this, perhaps because I recognize myself in the student, being in an inferior position. I do not see it as arguing against the teacher, when the reason for not wanting to take the veil off is due to religion and not out of spite.

Course objectives, something that I do not feel too knowledgeable about. – Yes, course objectives – something I definitively have to read and learn more about.

As to the student's and the teacher's behaviors, I feel that I do not have the proper knowledge base – after one semester at the teacher education program I think it would be remarkable if I could analyze both student and teacher behaviors, something that I hope we will return to during teacher education.

How the teacher's attitude affects the student – overlaps with teacher's behavior, see above.

Longer perspective – I want to make changes right away, and need to include thoughts on how to make a sustainable change for all students wearing a veil, not just this particular student.

The professional has written about authority, which I have not done at any length. – I need to be able to put myself in the teacher's shoes and not only the student's. I would like to learn more about leadership.

Concrete solutions /.../ as I feel I do not have enough knowledge in order to make recommendations – I think I need more experience, i.e. more field-based education, in order to know what works and what does not.

Figure A4c. An example of a student comparison with a professional answer (continued).

Appendix B. Scoring rubric for the "Interactive examination"

Assessment of:	Level	
	Acceptable	Excellent
Observation *Can you describe the situation?*	The description is not prejudiced. The description might contain assumptions not shown in the situation.	The description is not prejudiced.
	The description contains relevant details. The description contains details that are peripheral or that lack significance.	The description focuses on relevant details.
	The description contains the perspectives of all those directly involved in the situation.	
	The description can be understood by someone who has seen or experienced the situation.	The description can be understood by someone who has *not* seen or experienced the situation.

Assessment of:	Level	
	Acceptable	Excellent
Analysis *Can you make a statement and an interpretation of the problem?*	The analysis identifies a problem.	
	What causes the problem? The situation is also interpreted with the help of factors not shown in the situation.	The situation is interpreted with the help of factors not shown in the situation. The situation is interpreted with the help of course literature or other relevant sources.
	The analysis discusses conceivable motives for the behaviors shown.	The analysis discusses conceivable motives for the behaviors shown with the help of course literature or other relevant sources.
	The analysis discusses conceivable consequences of the situation.	The analysis discusses conceivable consequences of the situation with the help of course literature or other relevant sources.

Assessment of:	Level	
	Acceptable	Excellent
Taking action *Can you formulate actions to be taken, which take into consideration the needs of all those involved?*	Gives no suggestions as to what additional information is needed in order to make a decision.	Gives, when relevant, suggestions as to what additional information is needed in order to make a decision.
	Suggests actions that take both teacher's and students' needs into consideration.	Suggests several actions to be taken.
	Suggests action to be taken that takes only the immediate situation into consideration.	Suggests action/s to be taken that take/s both the immediate situation, as well as a longer perspective, into consideration.
	The actions suggested follow logically from the observation and the analysis.	

Assessment of:	Level	
	Acceptable	Excellent
Self-assessment *Can you use own and other's experiences as a basis for reflection and development?*	The comparison only identifies differences in *shape* or *quantity* between own and the other's analysis.	The comparison identifies differences in *subject, attitude* or *interpretation* between own and the other's analysis.
	The comparison contains reasons for the identified differences between own and the other's analysis.	The comparison argues for the own standpoints with the help of course literature or other relevant sources.
	The comparison identifies shortcomings in the competency as a teacher.	The comparison identifies shortcomings in the competency as a teacher, and formulates concrete learning needs.

Appendix C. Excerpts from the exemplars

This appendix shows short excerpts from the exemplars distributed to the students for the 2005 and 2006 versions of the "Interactive examination" for student teachers (Figure C1).

Assessment of: Observation

Student answer 1
The teacher is moving some furniture and the children are watching. Afterwards the teacher asks how many chairs there are on the right side of the basket. The teacher does not seem to comprehend that the children are not able to distinguish between right and left.

Aspect	Criter.	*N.Acc.*	*Acc.*	*Exc.*	*Comments*
Observation	1	✘			The description is prejudiced.
	2	✘			The description lacks several important details (which are then used to analyze the situation).
	3	✘			A description of the children's situation is lacking.
	4	✘			The description can not be understood by someone who has not seen or experienced the situation.

Abbreviations: Criter. = Criterion; N. Acc. = Not acceptable; Acc. = Acceptable; Exc. = Excellent.

Student answer 4

There are four children and a teacher in the room. In the beginning there are four chairs below the window with a box to the left and a basket to the right. The teacher tells the children that he is going to move the chairs a bit. The children become a bit worried and they say "move". When he has moved one chair to the right so that it is now on the right side of the basket, and out of the picture, he walks back to the children who are sitting in an irregular circle. Now there are three chairs below the window and at least one on the right side of the basket (out of picture). The children say "three" before the teacher has said anything. The teacher asks how many chairs there are on the right side of the basket. The children say "three" several times. The teacher asks them to listen to what he says and that they should raise their hands. He repeats the question. First, there is one child raising her hand and another answering "three". The children raise their hands and answer at the same time. Above all, there is one boy repeating "three" several times.

Aspect	Criter.	*N.Acc.*	*Acc.*	*Exc.*	*Comments*
Observation	1			✓	The description is not prejudiced.
	2		✓		The description contains details that are peripheral and/or lack importance.
	3		✓		The description includes both the teacher and the children.
	4			✓	The description can be understood by someone who has not seen or experienced the situation.

Abbreviations: Criter. = Criterion; N. Acc. = Not acceptable; Acc. = Acceptable; Exc. = Excellent.

Figure C1. The figure shows two student answers (of different quality) taken from the exemplars. The answers are assessed according to the scoring rubric and comments are given to each criterion.

Appendix D. References to papers from the Xpand project

Johnsson, Annette (2005). Lärarstudenters analys av sitt eget lärande och deltagande i grupparbete [Student teachers' analysis of their own learning and participation in group work]. *Rikskonferens för lärarutbildare i naturvetenskap – NALUT*, Malmö, Sweden.

Johnsson, Annette (2006). Lärarstudenters metaanalys av grupparbete [Student teachers' meta-analysis of group work]. *Netlearning Conference*, Ronneby, Sweden.

Johnsson, Annette (2006). What Impact does the level of dialogicality have on a discussion on the Net and its outcome? *Annual meeting of the Nordic Educational Research Association*, Örebro, Sweden.

Malmberg, Claes (2003). Samarbetslärande på nätet [Netbased collaboration]. In Anders Lindh & Bengt Linnér (Eds.), *Nära samarbete på distans* [Close collaboration at a distance] (Vol. 5). Malmö: Malmö University.

Malmberg, Claes (2006). *Kunskapsbygge på nätet. En studie av studenter i dialog* [Net-based knowledge building. A study of students in dialogue.]. Unpublished doctoral dissertation, Malmö University, Sweden.

Malmberg, Claes, Njord, Stefan, & Svingby, Gunilla (2005). Appropriation of cultural tools in an asynchronous computer mediated dialogues, *ISCAR Conference*, Sevilla, Spain.

Malmberg, Claes, Njord, Stefan, & Svingby, Gunilla (2005). Cultural tools as mediators in Computer Supported Collaborative Learning Environments, *ISCAR Conference,* Sevilla, Spain.

Malmberg, Claes & Svingby, Gunilla (2004). Students' communication and learning in computer supported dialogues. *Annual meeting of the Nordic Educational Research Association*, Reykjavik, Iceland.

Malmberg, Claes & Svingby, Gunilla (2004). Students' dialogues as contributions in education for sustainable development. In Per Wickenberg, Harriet Axelsson, Lena Fritzén, Gustav Helldén, & Johan Öhman (Eds.), *Learning to change our world*. Lund: Studentlitteratur.

Sandkull, Bengt & Svingby, Gunilla (2001). Accessibility and learning in higher education. In Pierre Dillenbourg, Anneke Eurelings, & Kai Hakkarainen (Eds.), *European perspectives on computer-supported collaborative learning.* Maastricht, NL: University of Maastricht.

Svingby, Gunilla (2001). Collaborative learning as the basis for flexible learning. *Application to the National Agency for Distance Education.* Malmö: Malmö University.

Svingby, Gunilla (2003). *Accessibility and Learning in Higher Education: Learning and New Media. Final report*. Malmö: Malmö University.

Svingby, Gunilla & Malmberg, Claes (2004). ALHE – a system to enhance accessibility by net based dialogues. *Annual meeting of the Nordic Educational Research Association*, Reykjavik, Iceland.