

Paper to be presented at the AEA-Europe Conference 2007, Stockholm, Sweden

The Use of Transparency in the “Interactive Examination” for Student Teachers

a.k.a.

The Use of Self-Assessment Criteria in the “Interactive Examination” for Student Teachers

Anders Jönsson

Malmö University, Sweden

* Correspondence concerning this paper should be addressed to Anders Jönsson, Malmö University, School of Teacher Education, SE-205 06 MALMÖ, Sweden.
E-mail: anders.jonsson@mah.se

Abstract

If the aim of education is for all students to learn and improve, then the expectations should be transparent to the students. In this study, three aspects of transparency is investigated: self-assessment criteria, a scoring rubric, and exemplars. The examinations studied were carried out in 2004, 2005 and 2006 respectively, all with a cohort of first year student teachers ($n = 170$, 154, and 138). There was a large difference in scores between the 2004 and 2005 cohorts ($d = 3.21$), when changes in the examination were implemented. The comparison between 2005 and 2006, when no changes were made, does not show a corresponding difference ($d = .27$). These results suggest that, by making the assessment more transparent, students' performances could be greatly improved.

The Use of Transparency

In recent research on assessment, norm-referenced testing has been shown to have some serious negative effects on students' learning, for instance by primarily assessing decontextualized knowledge, instead of competent and authentic performance, or by rewarding competition rather than collaboration. This has led to an extensive critique against "traditional testing" based on psychometric assumptions (Broadfoot, 1996; Gipps, 2001; Shepard, 2002), and in parallel to this critique, the interest in alternative modes of assessment has increased. These alternative modes of assessment are thought to:

- Support student learning – as opposed to only measuring it,
- Focus on more complex forms of knowledge – as opposed to atomized knowledge and rote learning,
- Motivate low-performing students – as opposed to the negative stress from constant failure, and
- Be fair in relation to different societal groups (ethnic, gender etc.) – as opposed to letting some particular group have a constant advantage over the others (Broadfoot, *ibid.*; Korp, 2003).

The above mentioned points are some of the cornerstones in what is often called the "assessment culture" or the "assessment paradigm" (see for example Birenbaum et al., 2006; Dochy, Gijbels, & Segers, 2006). The "assessment culture" is a rather new, and consequently a quite heterogenic, research paradigm. This is obvious when considering the many different areas of interest that are associated with the change in assessment culture, such as sociology (e.g. Broadfoot, *ibid.*), educational measurements, or "edumetrics" (e.g. Gielen, Dochy, & Dierick, 2003) and effects on student learning (e.g. Black & Wiliam, 1998). Still, there are some

fundamental issues that seem to be transcending the paradigm as a whole, and one such issue is the quite radical change in epistemology. While the psychometric tradition has been working mainly with decontextualized assessments of latent psychological traits, such as generic intelligence or other abilities, the change towards a constructivist, or even sociocultural, epistemology implies that these generic traits might not exist, and that all human knowledge is highly contextualized (Biggs, 1996.; Shepard, *ibid.*; Wertsch, 1991; 1998).

According to Shepard, the change in epistemology is one of major reasons that the “assessment culture” is gaining terrain. Another is the change in what perhaps could be called “assessment theory”. If working within the “assessment culture”, where the students are not to be compared to each other as is norm-referenced assessment, then the assessment has to be made in relation to some other point of reference, such as assessment criteria. Criterion-referenced assessment can be performed in many different ways and this concept has also been the starting point of much empirical research. In the “assessment culture”, however, one of the main reasons for the interest in criterion-referenced assessment is the transparency provided by such an approach.

In their influential article “A systems approach to educational testing”, Fredriksen and Collins (1989) put forth the concept of “transparency” as an important aspect of assessment validity. The terms of the assessment must be clear to the students, otherwise it has no potential for motivating and directing learning (see also Linn, Baker, & Dunbar, 1991). These researchers argue that the terms of the assessment should be so clear, that the students can assess themselves and their peers with the same accuracy as the teacher. One of the main problems here, however, is that students often have difficulties understanding the criteria (Orsmond & Merry, 1996). This could be remedied in several different ways, for example by letting the students work repeatedly

with the criteria in order to discover for themselves the proper notion of the criteria used. Although this method might be very time consuming, it is often seen as necessary in piaget-inspired constructivism, where discovery is more or less seen as the prerequisite of true understanding (Säljö, 2005). In a sociocultural perspective, on the other hand, this trail-and-error process would not be seen as appropriate, since in this view there are no "true" or abstract meanings of a criterion, which can be taken directly from a dictionary or a textbook. Instead criteria are seen as reflecting what is considered quality within a specific community of practice (Lave & Wenger, 1991), that is at a particular time and place. This means that the criteria are not selected at random, and that not all meanings of a criterion are of equal value. Thus it would make more sense to guide the students in the proper direction, which would be towards the specific meaning of the criteria as they are used in the particular community of practice that the students are aiming to become participants of. How then, can the students be guided towards understanding the criteria? This could surely be done in several different ways too, but recent research on assessment points towards some approaches of particular interest, which are inspired by the concept of transparency.

Self-Assessment

The first approach is pointed out by Fredriksen and Collins, and that is self-assessment. Self-assessment, just like criterion-referenced assessment, can be performed in many different ways and for many different reasons. The research on this topic is also influenced by the scientific paradigm within which it is conducted, and 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). Although quantitative

comparisons between students and teachers are not necessarily considered the most interesting aspect of self-assessment in the "assessment culture", since this comparison says very little about *what* the students and their teachers really agree (or disagree) on – as there are seldom any explicit criteria to relate the assessment to, there are some interesting findings when analyzing the studies at a meta-level. For the line of reasoning in this paper, one such finding is that the contextualized view on knowledge discussed above, is very much supported by this empirical research on self-assessment. 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, 1989). 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 that is also supported by 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.

In the most basic setting of criterion-referenced self-assessment, students would use the same criteria as the teacher to assess their own performance. But they also need to be assessed on their self-assessment skills, and be given feedback, if they are to learn and improve these skills as well. In a more advanced setting, the students could therefore use a second set of criteria, alongside the assessment criteria used by the teacher. These "criteria of self-assessment" could be used in order to make explicit how to self-assess in the specific context, and in this way further scaffold the process of self-assessment. The students could thus assess their performance with the aid of the "regular" assessment criteria, but they could also be shown how high quality

self-assessment is carried out within the specific community of practice. This design could potentially aid the students in developing their meta-cognitive skills, by making the self-assessment process more transparent.

Exemplars

The self-assessment criteria, however, are likely to be as difficult to understand as any other assessment criteria, and therefore the addition of such criteria alone are not likely to remedy the situation. The second approach is thus to provide the students with an array of exemplars assessed with, and followed by qualitative feedback in relation to, the criteria. This could show the students how the criteria are to be interpreted in the specific context and in this way make it clear to them what is expected of them, but also what is important to learn in the field of knowledge in which they are novices.

Scoring Rubrics

Besides the practice in self-assessment, Fredriksen and Collins (1989) suggest some other strategies in order to make the assessment transparent to the students, such as giving feedback on strengths and weaknesses and advice on how to improve. They also argue for providing “landmarks of success in performance” (p. 31), so that students have something to strive for. Even if criteria facilitate the search for qualities that can not be measured quantitatively, and also make it possible to estimate these qualities (see for example Eisner, 1991), they do not in themselves provide any “landmarks of success in performance”. There is, however, an assessment tool that combines criteria with different levels of quality performance, and that is the scoring rubric.

Even though the term “rubric” is used in several different ways, a widespread definition of the rubric states that it is a scoring tool for qualitative rating of authentic or complex student work. It includes criteria for rating important dimensions of performance, as well as levels of attainment for those criteria. The rubric tells both instructor and student what is considered important and what to look for when assessing (Arter & McTighe, 2001; Wiggins, 1998). By specifying criteria *and* different levels of quality performance, the rubric seems to be able to do some of the things that Fredriksen and Collins are asking for. This is also confirmed in a recent literature review on rubrics (Jonsson & Svingby, in press), where it appears to be easier for teachers to give feedback, as well as suggestions on how to improve performance, when the assessment is supported by a rubric. Furthermore, student self-assessment seems to be facilitated by the use of rubrics, and some studies show that students actually internalize the criteria, making them their own, and use them while self-assessing. There are also some indications that the use of rubrics might enhance learning. For example, in one study (by Sadler & Good, 2006), the students scored their own tests using a rubric, and the students improved their performance quite dramatically. Finally, in combination with exemplars, rubrics seem to enhance the consistency in scoring, which might be helpful if the students are to assess themselves and their peers with the same accuracy as the teacher.

In the study reported here, a combination of the above mentioned approaches has been used in order to increase the transparency in the assessment and to help students develop their self-assessment skills. This is done in a teacher education context, using an assessment methodology called the “Interactive examination”. The question guiding the study is whether this increased transparency, operationalized as the use of self-assessment, rubrics and exemplars, results in an improvement of student performance.

*Method**The "Interactive Examination"*

The "Interactive examination" is a structured assessment methodology originally developed and evaluated in the Faculty of Odontology at Malmö University, Sweden (Mattheos et al., 2004). In this methodology, assessment of self-assessment skills is included in a regular examination (just as suggested previously) and students' content specific skills and competencies are thus assessed in parallel to their self-assessment skills. Also, the methodology makes use of modern information- and communication technology in order to facilitate practice and feedback without necessarily increasing the workload of the personnel.

The "Interactive examination" has been adapted to, and implemented in, the teacher-education domain, where the examination is used to assess certain aspects of teacher competency, such as observational and analytical skills. The examination has so far been carried out for three years in a row, starting in 2004. Earlier research, conducted in relation to the examination, has shown that the "Interactive examination" indeed seems to be a promising way to assess complex performance in a valid manner (Jonsson & Baartman, 2006; Jonsson, Mattheos, Svingby, & Attström, 2007). The methodology consists of the following six stages:

Quantitative self-assessment, where the students estimate their own competency through a number of Likert-scale questions, graded from 1 (poor) to 6 (excellent). These self-assessment questions are equivalents to the criteria in the scoring rubric, making possible a comparison of students' self-assessment to their actual examination results. The purpose of this comparison is to highlight differences between the student's and the assessor's judgment, and not to constitute

a judgment per se (i.e. only for formative use). Therefore, possible deviations are only communicated to the students as a basis for reflection.

Analysis of critical situations, where the students, via the Internet, watch three short (1-2 minutes) movie sequences, showing different critical incidents in a classroom context. For example, there is one sequence where a Muslim girl is requested by the science teacher to remove her veil for safety reasons. Together with the movie sequences some background data is available, as well as the possibility to access the dialogue in text format. The allocated time to respond was approximately one hour per movie sequence. With each movie sequence as a starting point, the students respond to three different tasks, whereby they: (i) describe the situation, (ii) analyze the situation, and (iii) suggest how the teacher in the movie should proceed. These tasks are called "Observation", "Analysis", and "Taking action" respectively.

Expert comparison, where the students, after submitting their answers, receive conceivable ways of handling the same situations, written by experienced teachers, or "experts". This expert answer does not correspond to the best or the only solution, but rather to a justified rationale from an experienced colleague, which remains open to discussion. The expert documents have been written in advance and the students are given access to them as they submit their responses to the personal tasks. This is a way of dealing with the problem of providing timely feedback to a large number of students, but the expert answers also provide a kind of social interaction, although in a fixed (or "frozen") form.

By the aid of the expert answer, the students can, according to the concept of "the zone of proximal development" (Vygotsky, 1978), potentially reach further than they can on their own, thus making the assessment *dynamic*. Dynamic assessment means that interaction can take place, and feedback can be given, during the assessment or examination, which separates it from more

”traditional assessments” (Swanson & Lussier, 2001). In this way, dynamic assessment provides the possibility to learn from the assessment, but also to assess the student’s potential (”best performance”), rather than (or together with) his or her ”typical performance” (Gipps, 2001). Meta-analyses of empirical studies has shown that dynamic assessment indeed seem help to improve student performance, and also that low-performing students are those who benefit the most, thus making the difference between high- and low-performing students less pronounced (Swanson & Lussier, *ibid.*).

After receiving the expert document, the students must, within a week, prepare a comparison document. In this document they compare their own answer to that of the experienced teacher, and identify differences between their own and the expert answer. The students are also expected to reflect on the reasons for these differences and try to identify own needs for further learning. This comparison document is a qualitative form of self-assessment in the “Interactive examination”, and it is supported by “self-assessment criteria” in the rubric, making it explicit how high quality self-assessment is expected to be carried out in this particular context. In contrast to the quantitative self-assessment, the qualitative self-assessment is used for summative purposes as well.

Evaluation following the examination, where the students evaluate the experience. Most of the questions are of Likert-type on a scale from 1 to 9. This stage is not a part of the examination, as the students are not assessed on the basis of their evaluation results, but the evaluation is an integrated part of the examination methodology as described by Mattheos et al. (2004). Also, this evaluation is used for research purposes, investigating for instance students’ perceptions of the examination.

Assessment, where the students' performances, that is (i) their analyses of the critical situations and (ii) their comparisons with expert answers, are assessed by an external assessor (i.e. not one of the instructors). As was mentioned previously, a scoring rubric has been developed to aid this assessment procedure. The rubric is thought to facilitate more reliable scoring, but also to communicate to the students what is expected of them. Therefore, the students had access to the rubric well before the examination.

The assessment of students' performances was made in relation to the criteria in the rubric, which consisted of 15 criteria (16 for year 2004) with three levels of quality (Fail, Acceptable, and Excellent). As each student analyzed three different movie sequences, the levels were transformed into marks (0, 1, 2 marks) in the assessment procedure, which means that the students received a score somewhere from 0 to 6 for each criterion. This is the same interval as the Likert scales in the initial quantitative self-assessment, making possible a comparison between the self-assessment and the examination score. For research purposes, an overall score has been computed from the sub-scores for the separate criteria, for example in order to estimate the interrater reliability, but no overall scores have been communicated to the students. This since no other grades other than "Pass" or "Fail" has currently been used in the course in question.

Feedback was provided on student performance in relation to each criterion in the rubric, and if a student failed a specific criterion for all three movie sequences, he/she had to rewrite his/her analysis or comparison document until an acceptable level of quality was reached.

Changes in the “Interactive Examination”

The “Interactive examination” was carried out for the first time in 2004, after which a thorough analysis was made of students’ answers and results. This analysis resulted in three major changes, which were implemented in the 2005 version of the examination and they will be described in more detail below. No further changes were made between the 2005 and 2006 versions of the examination.

After the 2004 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 the criteria which a large portion of the students did not fulfill, or only barely fulfilled, were indistinctly expressed – and not that these criteria necessarily were more difficult than the others or that those students who succeeded had to be more gifted than the others. This analysis led to a reformulation of some of the criteria in the rubric. For example, the first criterion in the rubric, assessing students’ observational skills, was originally expressed only as “The description is not prejudiced”, but 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 was finding out which levels in the rubric that were not, or seldom, used by the assessor, since this could indicate that he had 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 discriminating between the levels most likely were due to the design (i.e. that the tasks or the methodology did not give the students 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 means that the 2004 version

of the rubric had 16 criteria, while the 2005 and 2006 versions only had 15. As the questions in the initial quantitative self-assessment are equivalents of the criteria in the rubric, these questions were also changed accordingly.

In connection to the examination in 2004, a qualitative analysis was made of the comparison documents written by the students. The results from this analysis was compared to a similar examination for dental students (Jonsson et al., 2007), 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 "expert's", or even better. There are many possible explanations for these results, but one hypothesis could be that the expert documents were not perceived as sufficiently professional by the student teachers. Thus 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 the student comparisons, for example which kinds of differences the students identified between their own answer and the "expert's" (Jonsson et al., 2007). This categorization also 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 could be 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 was used to compile a document with answers assessed and commented upon in relation to the rubric. Aiming to help the students interpret the criteria, this document was distributed to the students before the examination 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 were needed; (3) one criterion was removed; (4) those 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 expert documents were enhanced. Furthermore, the students in cohorts 2005 and 2006 could (6) access a document with assessed student answers, or exemplars.

Sample

The examinations providing data for this study 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 ($n = 170, 154, \text{ and } 138$). The students were exposed to the "Interactive examination" for the first time. Data have been collected at several occasions and all parts were not completed by all students. Therefore, the actual number of students in different parts of the analysis may differ somewhat.

Research data and analyses

Background data. Earlier research on the “Interactive examination” for the 2004 and 2005 cohorts have shown that there are no background variables (such as sex, parents’ education, ethnicity, computer experience, and alignment towards teaching earlier or later years of schooling) contributing significantly to the prediction of examination results when performing regression analyses (Jonsson et al., 2007; Jonsson & Baartman, 2006). The only exception was subject major, where geography majors performed significantly lower in 2005 as compared to mathematics and science majors. To make sure that there is no systematic disadvantage for geography majors, but also, since there might be slight variations in the relative frequency of students with different majors from year to year, to make sure that this does not significantly influence the results when comparing the different cohorts, students with different majors were compared using an analysis of variance (ANOVA) for the 2006 cohort.

Students’ examination scores. The students were assessed on their analyses of critical classroom situations as well as on their comparison with the expert analyses of the same situations. The examination results for the three cohorts were compared using an analysis of variance (ANOVA) and effect sizes were calculated. This comparison is done separately for the different tasks (“Observation”, “Analysis”, “Taking action”, and “Comparison with ‘expert’”), since the tasks might not be of equal difficulty. Also, the tasks are affected somewhat differently by the changes implemented between 2004 and 2005.

As can be seen in Table 1, two of the tasks are more or less directly comparable (“Observation” and “Comparison”), whereas the other two differ to some extent. In the “Analysis” task there is one more criterion in 2004 than in 2005/2006. This problem has been solved by comparing the mean score per criterion. However, the fact remains that one criterion is

merged differently, and the same thing occurs in “Taking action”. For example, if the “Acceptable” and “Excellent” levels have been merged, then the student can either get 0 or 2 marks (but not 1) for that criterion, but if “Not acceptable” and “Acceptable” have been merged the student can get either 1 or 2 marks (but not 0). Even though the maximum score is the same, this difference could potentially effect the distribution of marks, making students results on the task appear either higher or lower. This effect is likely to be very small, however, since the reason for merging the levels in the first place was that they were not used by the assessor in the 2004 version of the examination.

Table 1.

Characteristics for the 2004 and 2005 versions of the rubric. The 2006 version was identical to the 2005 Version.

Rubric characteristics	Task							
	Observation		Analysis		Taking action		Comparison	
	2004	2005	2004	2005	2004	2005	2004	2005
N:o criteria	4	4	5	4	4	4	3	3
Merged criteria ¹	0/2	0/2	1/2	0/2	1/2	0/2 1/2	-	-
Max. score	24	24	30	24	24	24	18	18

¹If the “Acceptable” and “Excellent” levels have been merged, then the student can either get 0 or 2 marks (but not 1), which is shown in the table as 0/2. If “Not acceptable” and “Acceptable” have been merged, this is shown as 1/2.

It should be noted that the assessor is held constant in the comparisons, even though the interrater reliability is acceptably high (i.e. Pearson's correlation above .9 for the overall score). In the 2005 version of the examination, two independent assessors were used (Assessor 1 and Assessor 2), while the cohorts of 2004 and 2006 were assessed by one assessor only (Assessor 1 and Assessor 2 respectively). Thus Assessor 1 is used in the comparison of 2004 and 2005, and Assessor 2 for the comparison of 2005 and 2006.

Student evaluation questionnaires. The students were asked whether they had used the rubric for the examination and, in 2005 and 2006, also in what way. Their answers were categorized as (in 2004) as "Used" or "Not used" and (in 2005 and 2006) as either "Active use" for those who had used the rubric during the examination, "Read only" for those who had only read through the rubric before the examination¹ or "Not used" for those answering that they have never heard anything about a rubric. Then a comparison of the examination results from these three groups (i.e. within each cohort) was made with an analysis of variance (ANOVA) and effect sizes were calculated.

It should be emphasized that this analysis is based on students' own reports on their rubric use in the evaluation questionnaire, and that experimental conditions has not been used. The main reason for this is the ambition to give all students equal and as optimal conditions as possible. Also, it could be considered unethical to provide only some students with a scoring rubric in an examination used for summative purposes, since then they would not be "competing" on the same terms. As the students are not selected at random, potential differences between students using and not using the rubric should be interpreted with care. For example, it

¹ Those students not specifying how they have used the rubric were also placed in this category.

might be that the more ambitious students were also the ones using the rubric, and perhaps these students should have performed better even without the rubric.

Results

The analysis of examination scores from students with different subject majors show no significant difference between the groups investigated.

The results from the comparison of students' scores for the three cohorts are shown in Table 2. As can be seen, there is a large difference between the 2004 and 2005 cohorts for all tasks, which is when the changes were implemented. For example, in the "Taking action" task the students in 2005 more than doubled their scores as compared to the students in 2004. The comparison between 2005 and 2006, when no changes were made, does not show a corresponding difference in student scores. Instead these differences are quite small, and sometimes not statistically significant despite the quite large sample.

Another thing that can be seen in Table 2, is that the tasks are not of equal difficulty. For instance, the mean criterion score for the observational task is persistently higher than for the other tasks. In the other end of the spectrum is the comparison task, which seems to be the most difficult. Even though "Taking action" and "Comparison" were of approximately equal difficulty in 2004, the increase in score in 2005 is not of equal magnitude for the comparison task.

Table 2.

Differences in students' results between the cohorts of 2004 and 2005, and between 2005 and 2006. Scores are presented as the mean score per criterion, since there were 16 criteria in 2004, but only 15 in 2005 and 2006. Effect sizes are reported as Cohen's d (Cohen, 1988).

Task	Mean score ¹ (SD)		Difference (%)	Effect size
Observation				
2004/2005	3.67 (.764)	5.40 (.483)	47.0	2.70***
2005/2006	5.22 (.653)	5.69 (.540)	9.1	.79***
Analysis				
2004/2005	2.58 (.693)	3.53 (.744)	36.5	1.31***
2005/2006	3.48 (.664)	3.72 (.692)	7.0	.36*
Taking action				
2004/2005	2.00 (.527)	4.49 (.640)	124.7	4.25***
2005/2006	4.52 (.604)	4.68 (.948)	3.5	.20
Comparison				
2004/2005	2.12 (.903)	3.23 (1.139)	52.0	1.07***
2005/2006	3.46 (1.155)	3.18 (.882)	- 7.9	.27*
Overall score				
2004/2005	2.62 (.476)	4.25 (.535)	62.0	3.21***
2005/2006	4.23 (.542)	4.38 (.563)	3.6	.27*

¹Note: Range for all criteria is 0-6.

* p < .05; *** p<.001

The results from the analysis of rubric usage are shown in Table 3. As can be seen, those who state that they have used the rubric have significantly higher scores in 2004 and 2005. This difference is even somewhat larger if “Active use” is compared to the other categories of rubric use in 2005. Although the same trend is seen in all three cohorts, in 2006 there is no statistically significant difference in scores between those who used the rubric and those who did not, and only a small effect of active use versus other forms of rubric use (and non-use). As could be seen in the table, this is due to the higher mean score of the students reporting that they have not used the rubric in 2006 as compared to 2005. It should be noted that not all students have answered this question in the questionnaire, and that there is no significant difference between the scores of those who have answered the question and those who have not.

Discussion

The question guiding this study was whether the increased transparency, operationalized as the use of self-assessment, rubrics and exemplars, result in an improvement of student performance. Although the effect of these three aspects of transparency can not be separated in the current research design, they will be discussed individually in order to highlight interesting results.

Table 3.

Differences in students' results between different categories of rubric usage. The effect size is reported as Cohen's d (Cohen, 1988).

Categories of rubric use and effect sizes	Cohort					
	2004		2005		2006	
	Mean score ¹ (SD)	n	Mean score (SD)	n	Mean score (SD)	n
Not used	40.0 (6.965)	76	57.4 (7.645)	23	64.5 (6.015)	25
Read only	-	-	63.8 (7.443)	63	64.6 (8.866)	40
Active use	45.12 (7.603)	65	66.9 (6.798)	43	68.1 (7.544)	34
Effect size:						
Used vs. Not used	.707***		1.02***		-	
Active use vs. others	-		1.30***		.454*	

¹Note: Maximum score in 2004 was 96, while in 2005 and 2006 it was 90.

* p < .05; *** p < .001

The Use of Self-Assessment

In the “Interactive examination” students self-assess both quantitatively and qualitatively.

The quantitative self-assessment is only used for formative purposes, and the comparison of

students' own judgment with the examination results can only occur after the examination is finished. As a consequence, this self-assessment is not thought to affect the examination scores in any significant way. It could, however, have a potential effect if the students were confronted with the same assessment methodology again.

In the qualitative self-assessment (i.e. the comparison document) the students are not only prompted to self-assess their performance with the assessment criteria in the rubric, but they are also guided in the process by a set of "criteria of self-assessment". This is thought to scaffold the process of self-assessment and make it more explicit, since self-assessment is many times described as a cognitively very complex task (Topping, 2003). The student scores on the qualitative self-assessment part of the examination supports this view, as this indeed seems to be the most challenging part of the examination. Although the students had access to criteria to guide their self-assessment, along with exemplars, the mean score for the 2005 and 2006 cohorts are only about half the maximum score. This could be compared to the "Taking action" task, where the corresponding score is about three quarters of the maximum score.

As all three cohorts have had access to the "criteria of self-assessment", it is not possible to estimate the individual effect of these criteria. There is a large increase in scores for the comparison task from 2004 to 2005, when the criteria in the rubric were clarified, but this change is implemented at the same time as the exemplars, which may account for some (or all) of this effect. The self-assessment criteria are, however, part of the rubric, and this aspect will be discussed next.

The Use of a Rubric

The rubric is thought to tell both instructor and student what is considered important in the assessment. By specifying criteria, as well as different levels of quality performance, the rubric could make it easier for teachers to give feedback and suggestions on how to improve performance. It could also facilitate student self-assessment.

The results from this study indicate that those students using the rubric actively perform significantly better than those who only read through the rubric, or those who do not use it at all. Although this statement can not be unequivocally proved by the current research design (i.e. non-experimental settings and that not all students have answered this question in the questionnaire), the difference is quite large for those who used, versus those who did not use, the rubric in 2004 and 2005, and even larger if “Active use” is compared to the other categories. This trend is evident in all three cohorts, even though not as strong in the 2006 cohort, and it would therefore seem reasonable to think that at least some of this effect is attributable to the use of the rubric.

Some of the criteria in the rubric were clarified and more distinctly expressed in the 2005 version of the rubric, as compared to the 2004 version. A possible effect of this change is seen when comparing the effect sizes for these two cohorts, where the “rubric effect” is larger in 2005 than in 2004, implying that it is indeed this reformulation of criteria that contributed to this change. There is, however, not such a large effect in the cohort of 2006, since those students reporting that they have not used the rubric has a higher mean score as compared to the corresponding students in 2005. A possible explanation for this might be that a larger group of students have used neither the rubric nor the exemplars in 2005, thus getting lower scores, while in 2006 most students have used the exemplars even if they have not used the rubric. This

explanation might also account for the fact that students' results were somewhat higher for all tasks in 2006, except for the comparison task where the use of the rubric, with the self-assessment criteria, is perhaps most important. Unfortunately, there are currently no data available on students' use of the exemplars to verify this hypothesis.

The Use of Exemplars

Criteria are difficult to understand, especially for novices, and the students should therefore ideally be provided with exemplars assessed with, and followed by qualitative feedback in relation to, these criteria. This could show the students how the criteria are to be interpreted in the specific context and in this way make it clear to them what is expected of them.

In this study, exemplars were provided for the cohorts of 2005 and 2006, but not to the cohort of 2004. When comparing student scores from these years, a very large increase can be observed between those who had access to the exemplars and those who did not. As the effect due to clarifications in the rubric can not be used to explain the full magnitude of this increase, it would seem that the use of exemplars could have a large impact on student performance (at least in combination with self-assessment and a rubric). Before this conclusion is drawn, however, there is still one change in the assessment methodology that has not yet been discussed, and that is the thoroughness and professional appearance of the expert documents.

The Expert Documents

By changes made in the expert documents to become more thorough and systematic, it could have become easier for the students to learn how to write their own answers, and it could also have become easier to do their comparison with the "expert", both of which could

potentially affect the results on the “Interactive examination”. The increase in scores between the cohorts of 2004 and 2005 could therefore, at least partly, be attributed to the changes in the expert documents. There are, however, two major counter arguments that contradict this assumption:

1. If it had become easier to do the comparison with the “expert”, then the scores for this task would have increased. As is shown in Table 2, it is not only the scores on this task that increases, however, but the scores on the other tasks increase as well. Also, the increase is larger in some of the other tasks as compared to the qualitative self-assessment (e.g. “Taking action”). Students’ scores on the “Observation” task does not increase quite as much as on the “Comparison” task, but this is more likely to be a result of ceiling effects, since students generally had a much higher scores on the “Observation” task than on the others (which might be expected, as describing a situation is probably easier than analyzing it or coming up with an action plan).
2. If it had become easier for the students to learn how to write their own answers, by the aid of the expert documents, than there would be a large difference between their first answer (where they had not yet gained access to any of the expert documents) and the two following ones (where they had access to at least one expert document). Such a difference can not be observed, either for individual students or for the students as a group. In fact the students are generally very consistent in their performance (Cronbach’s $\alpha = .866$). A major reason for not learning from the expert documents during the examination, is probably the time frame, as there is no time for the students to take a closer look at the expert documents at the same time as they are writing their own analyses.

Conclusions: The Use of Transparency

The results from this study suggest that, by making the assessment more transparent, students' performances could be greatly improved. This is in line with the sociocultural epistemology outlined in the introduction, where context dependent skills (such as self-assessment) are not generic qualities we are born with, but things that are given meaning in a particular context and that can be learned by appropriate instruction – and assessment. These results also point towards the importance of communication in instruction, since transparency is really a way of communicating to the students what is important and what is expected of them. And if the true aim for instruction is for all students to learn and improve, then increased transparency by the use of self-assessment, exemplars, and rubrics seems to help the students understand what is expected of them and to improve their performance.

Unfortunately, there are limitations in the current research design, making it impossible to estimate the separate effects of the different aspects of transparency in this study, and future research could perhaps try to “split them apart” in order to investigate their individual contributions more closely. But this has in part already been done, as there are several empirical studies reporting on for instance exemplars and rubrics (see Jonsson & Svingby, in press), and according to the epistemological stance taken by the “assessment culture”, there is really no such thing as a decontextualized assessment. Perhaps more interesting then, should be for future research to implement and optimize the different aspects in many diverse contexts, since it unlikely that neither their individual contributions, nor their collective effect, will be the same in different settings.

The limitations discussed above stem from the fact that the research is carried out in line with the assumptions of the “assessment culture”. As a consequence, the research is done in natural settings, where the ambition has been to give all students equal and as optimal conditions as possible to succeed. Thus it has not been feasible to provide only some students with self-assessment criteria, exemplars, or a rubric in order to compare their results. There are, however, also strengths to this design. For example, as the research is carried out in a natural setting and the assessment is used for summative purposes, students are likely to invest a lot of effort into the assessment, which might not always be the case when doing research in settings where the students have no real interest in performing well. And last, but not least, the comparison of scores *between* the cohorts are not affected by the limitations in the research design to the same extent as the analysis of rubric use, since the conditions of who had access to the exemplars and who had not, were controlled. So even though the individual effects of the different aspects could not be separated in this study, the comparison of scores between the cohorts of 2004 and 2005 can be used as an estimation of the combined effect, which is quite large (3.21 for the overall score). This could be compared to other effect sizes reported on formative assessment, which are usually between .4 and .7 (Black & Wiliam, 1998)

References

- Arter, J. & McTighe, J. (2001). *Scoring rubrics in the classroom*. Thousand Oaks: Corwin Press.
- Biggs, J. (1996). Enhancing teaching through constructive alignment. *Higher Education*, 32, 347-364.
- Birenbaum, M., Breuer, K., Cascallar, E., Dochy, F., Dori, Y., Ridgeway, J., et al. (2006). A learning integrated assessment system. *Educational Research Review*, 1, 61–67.
- Black, P. & Wiliam, D. (1998). Assessment and classroom learning. *Assessment in Education*, 5, 7-74.
- Boud, D. & Falchikov, N. (1989). Quantitative studies of self-assessment in higher education: a critical analysis of findings. *Higher Education*, 18, 529-549.
- Broadfoot, P. (1996). *Education, assessment and society – a sociological analysis*. Buckingham: Open University Press.
- Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. Hillsdale: L. Erlbaum Associates.
- Dochy, F., Gijbels, D., & Segers, M. (2006). Learning and the emerging new assessment culture. In L. Verschaffel, F. Dochy, M. Boekaerts, & S. Vosniadou (Eds.), *Instructional psychology: Past, present and future trends*. Oxford, Amsterdam: Elsevier.
- Dochy, F., Segers, M., & Sluijsmans, D. (1999). The use of self-, peer and co-assessment in higher education: a review. *Studies in Higher Education*, 24, 331-350.
- Eisner, E. (1991). Taking a Second Look: Educational Connoisseurship Revisited. In M. W. McLaughlin & D. C. Phillips (Eds.), *Yearbook of the National Society for the Study of Education. Evaluation and education at quarter century* (p. 169 – 187). Chicago: The National Society for the Study of Education.

- Falchikov, N. & Boud, D. (1989). Student self-assessment in higher education: A meta-analysis. *Review of Educational Research*, 59, 395-430.
- Frederiksen, J. R. & Collins, A. (1989). A systems approach to educational testing. *Educational Researcher*, 18, 27-32.
- Gielen, S., Dochy, F., & Dierick, S. (2003). Evaluating the consequential validity of new modes of assessment: The influence of assessment on learning, including pre-, post-, and true assessment effects. In M. Segers, F. Dochy & E. Cascallar (Eds.), *Optimizing new modes of assessment: In search of qualities and standards* (p. 37-54). Dordrecht: Kluwer Academic Publishers.
- Gipps, C. (2001). Sociocultural aspects of assessment. In G. Svingby & S. Svingby (Eds.), *Bedömning av kunskap och kompetens* [Assessment of knowledge and competence] (p. 15-67). Stockholm: Lärarhögskolan i Stockholm, PRIM-gruppen.
- Jonsson, A. & Baartman, L. K.J. (2006). Estimating the quality of new modes of assessment: The case of an “Interactive Examination” for teacher competency. *EARLI SIG Assessment Conference 2006*, Northumbria, UK.
- Jonsson, A., Mattheos, N., Svingby, G., & Attström, R. (in press). Dynamic assessment and the “Interactive examination”. *Educational Technology & Society*.
- Jonsson, A. & Svingby, G. (in press). The use of scoring rubrics: Reliability, validity and educational consequences. *Educational Research Review*.
- Korp, H. (2003). *Kunskapsbedömning – hur, vad och varför* [Assessment of knowledge – how, what, and why]. Stockholm: Myndigheten för skolutveckling.
- Lave, J. & Wenger, E. (1991). *Situated learning: legitimate peripheral participation*. Cambridge: Cambridge University Press.

- Linn, R. L., Baker, E. L., & Dunbar, S. B. (1991). Complex, performance-based assessment: Expectations and validation criteria. *Educational Researcher*, 20, 15-21.
- Mattheos, N., Nattestad, A. Falk Nilsson, E., & Attström, R. (2004). The interactive examination: Assessing students' self-assessment ability. *Medical Education*, 38, 378-389.
- Orsmond, P. & Merry, S. (1996). The importance of marking criteria in the use of peer assessment. *Assessment & Evaluation in Higher Education*, 21, 239-250.
- Sadler, P. M. & Good, E. (2006). The impact of self- and peer-grading on student learning. *Educational Assessment*, 11, 1-31.
- Shepard, L. A. (2002). The role of classroom assessment in teaching and learning. In V. Richardson (Eds.), *Handbook of Research on Teaching*, 4th ed. (p. 1066-1101). Washington DC: American Educational Research Association.
- Swanson, H. L. & Lussier, C. M. (2001). A selective synthesis of the experimental literature on dynamic assessment. *Review of Educational Research*, 71, 321-363.
- Säljö, R. (2005). *Lärande och kulturella redskap: om lärprocesser och det kollektiva minnet* [Learning and cultural tools: About learning processes and the collective mind]. Stockholm: Norstedts akademiska förlag.
- Topping, K. (2003). Self and peer assessment in school and university: Reliability, validity and utility. In M. Segers, F. Dochy & E. Cascallar (Eds.), *Optimizing new modes of assessment: In search of qualities and standards* (p. 55-87), Dordrecht: Kluwer Academic Publishers.
- Vygotsky, L. S. (1978). *Mind in society*. Cambridge: Harvard University Press.
- Wertsch, J. V. (1991). *Voices of the mind: a sociocultural approach to mediated action*. London: Harvester Wheatsheaf.

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

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