Available online at www.sciencedirect.com

ScienceDirect

Educational Research Review 2 (2007) 130–144

Educational Research Review

www.elsevier.com/locate/EDUREV

Research review

The use of scoring rubrics: Reliability, validity and educational consequences

Anders Jonsson [*], Gunilla Svingby

School of Teacher Education, Malmo University, SE-205 06 Malmo, Sweden

Received 3 August 2006; received in revised form 3 May 2007; accepted 4 May 2007

Abstract

Several benefits of using scoring rubrics in performance assessments have been proposed, such as increased consistency of scoring, the possibility to facilitate valid judgment of complex competencies, and promotion of learning. This paper investigates whether evidence for these claims can be found in the research literature. Several databases were searched for empirical research on rubrics, resulting in a total of 75 studies relevant for this review. Conclusions are that: (1) the reliable scoring of performance assessments can be enhanced by the use of rubrics, especially if they are analytic, topic-specific, and complemented with exemplars and/or rater training; (2) rubrics do not facilitate valid judgment of performance assessments per se. However, valid assessment could be facilitated by using a more comprehensive framework of validity when validating the rubric; (3) rubrics seem to have the potential of promoting learning and/or improve instruction. The main reason for this potential lies in the fact that rubrics make expectations and criteria explicit, which also facilitates feedback and self-assessment.

Keywords: Alternative assessment; Performance assessment; Scoring rubrics; Reliability; Validity

Contents

[*] Corresponding author.
E-mail address: anders.jonsson@lut.mah.se (A. Jonsson).

1747-938X/$ – see front matter © 2007 Elsevier Ltd. All rights reserved.
doi:10.1016/j.edurev.2007.05.002

1. Introduction

This article reviews studies that deal with the problem of assessing complex competences in a credible way. Even though the meaning of “credibility” can vary in different situations and for different assessment purposes, the use of scoring rubrics is increasingly seen as a means to solve this problem.

Today the assessment in higher education is going through a shift from traditional testing of knowledge towards “assessment for learning” (Dochy, Gijbels, & Segers, 2006). The new assessment culture aims at assessing higher order thinking processes and competences instead of factual knowledge and lower level cognitive skills, which has led to a strong interest in various types of performance assessments. This is due to the belief that open-ended tasks are needed in order to elicit students’ higher order thinking.

Performance assessment can be positioned in the far end of the continuum representing allowed openness of student responses, as opposed to multiple-choice assessment (Messick, 1996). According to Black (1998), performance assessment deals with “activities which can be direct models of the reality” (p. 87), and some authors write about *authentic assessment* and tasks relating to the “real world”. The notion of reality is not a way of escaping the fact that all learning is a product of the context in which it occurs, but rather to try to better reflect the complexity of the real world and provide more valid data about student competence (Darling-Hammond & Snyder, 2000). As a consequence, performance assessments are designed to capture more elusive aspects of learning by letting the students solve realistic or authentic problems.

When introducing performance assessment, the problem of whether observations of complex behaviour can be carried out in a credible and trustworthy manner shows up. This problem is most pressing for high-stakes assessment, and 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 assessment aiming to aid student learning is less influenced by this call for high levels of reliability but the assessment still needs to be valid. Since performance tasks are often assessed with the guidance of scoring rubrics, the effective design, understanding, and competent use of rubrics is crucial, no matter if they are used for high-stake or classroom assessments—although the primary focus of these two perspectives will differ.

From the perspective of high-stakes assessment, Stemler (2004) argues that there are three main approaches to determine the accuracy and consistency of scoring. These are consensus estimates, measuring the degree to which markers give the same score to the same performance; consistency estimates, measuring the correlation of scores among raters; measurement estimates, measuring for instance the degree to which scores can be attributed to common scoring rather than to error components.

It seems to be more difficult to state what should be required for assessments with formative purposes, as well as for combinations of formative and summative assessments. Nonetheless, most educators and researchers seem to accept that the use of rubrics add to the quality of the assessment. For example, Perlman (2003) argues that performance assessment consists of two parts: “a task and a set of scoring criteria or a scoring rubric” (p. 497). The term “rubric”, however, is used in several different ways: “there is perhaps no word or phrase more confusing than the term ‘rubric’. In the educational literature and among the teaching and learning practitioners, the word ‘rubric’ is understood generally to connote a simple assessment tool that describes levels of performance on a particular task and is used to assess outcomes in a variety of performance-based contexts from kindergarten through college (K-16) education” (Hafner & Hafner, 2003, p. 1509).

A widespread definition of the educational 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 standards 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; Busching, 1998; Perlman, 2003). This holds for both high-stake assessments and assessment for learning.

Two main categories of rubrics may be distinguished: holistic and analytical. In holistic scoring, the rater makes an overall judgment about the quality of performance, while in analytic scoring, the rater assigns a score to each of the dimensions being assessed in the task. Holistic scoring is usually used for large-scale assessment because it is assumed to be easy, cheap and accurate. Analytical scoring is useful in the classroom since the results can help teachers and students identify students' strengths and learning needs. Furthermore, rubrics can be classified as task specific or generic.

There are several benefits of using rubrics stated 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 raters. 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 like rubrics offer a way to provide the desired validity in assessing complex competences, without sacrificing the need for reliability (Morrison & Ross, 1998; Wiggins, 1998). Another important effect of rubric use often heard in the common debate, is the promotion of learning. This potential effect is focused in research on formative, self-, and peer assessment, but also frequently mentioned in studies on summative assessment. It is assumed that the explicitness of criteria and standards are fundamental in providing the students with quality feedback, and rubrics can in this way promote student learning (Arter & McTighe, 2001; Wiggins, 1998).

Still, there seem to be little information in the literature on the effectiveness of rubrics, when used by students to assess their own performance. Orsmond and Merry (1996) argue that students might not find the qualities in their work even if they know what to look for, since they have a less developed sense of how to interpret criteria. Differences between instructor and student judgments might thus well be attributed to the students' lesser understanding of the criteria used and not to the performance as such. It is therefore argued that rubrics should be complemented with "anchors", or examples, to illustrate the various levels of attainment. The anchors may be written descriptions or, even better, actual work samples (Busching, 1998; Perlman, 2003; Wiggins, 1998).

Even if the use of rubrics is gaining terrain, the utility may be limited by the quality of the scoring rubrics employed to evaluate students' performance. And even though the above mentioned benefits of rubrics may seem plausible, research evidence to back them up is needed, which is not always the case when the usage of rubrics is argued for in the common debate. This paper aims to investigate whether evidence can be found in the research literature on the effects of rubrics in high-stake summative, as well as in formative, assessment. The paper will try 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?

2. Procedure and data

Research on rubrics for assessing performance was originally searched online in the Educational Resources Information Center (ERIC), PsychINFO, Web of Science, and in several other databases, such as ScienceDirect, Academic Search Elite/EBSCO, JSTOR and Blackwell Synergy, complemented with search in Google Scholar and various reference lists. The search for rubrics/educational rubrics/scoring rubrics gave thousands of hits, which illustrates that the word is embedded in the vocabulary of teachers and teacher educators. The rubric seems to be a popular topic in the educational literature, and at educational conferences, which is seen by the body of literature that has accumulated in the past decade on design, construction, rationale, and use of rubrics as a tool for assessment of performance (Hafner & Hafner, 2003).

The search was then narrowed down to include only peer-reviewed articles in journals, conference papers, research reports and dissertations. No time limit was set. Only studies explicitly reporting on empirical research where rubrics were used for performance assessment were included, excluding a vast amount of articles on the development of rubrics, opinion papers on the benefits of rubrics, narratives from schools or colleges, and guides on how to use rubrics. Also, research that deals with other types of criteria or scripts for assessment has been excluded. This reduced the total number of papers included to 75.

Of the total number of studies reviewed, the majority was published during the last decade. Only seven articles were published before 1997. The distribution indicates that the rubric is a quite recent research issue. This notion is strengthened by the fact that the studies are found in 40 different journals, and only a handful of these have published more than one study on the subject. The variety of journals, on the other hand, from Applied Measurement in Education, and Educational Assessment over Assessing writing, and International Journal of Science Education to

Academic Medicine, and Bio Science, indicate the great educational interest in rubrics. Content, focus, type of rubrics used, as well as the actors involved, also vary considerably. The whole range from K-12, college, and university to active professionals is represented. Almost half of the studies focus on students and active professionals – many of which are student teachers and teachers – while the youngest children are less represented. Most frequent are studies on the assessment of complex teacher and teaching qualities, alongside with studies on writing and literacy. The types of performances studied represent a wide variety of competences, like critical thinking, classroom practice, engineering scenarios, essay writing, etc. The variation of the research field also shows itself in the research focus. The majority of studies are mainly interested in evaluating rubrics as part of an assessment system. Many of these studies focus on assessing the reliability merits of a specific rubric. Another large group consists of studies interested in making teachers' assessments more reliable with rubrics. About one-fifth of the reviewed studies have formative assessment in focus. Among those, studies are found that turn their attention to self- and peer assessment. These studies are, however, more frequently reported in the last few years, possibly indicating a growing interest.

The selected articles have been analyzed according to their research and rubric characteristics. Relevant characteristics were mainly educational setting (e.g. elementary, secondary or tertiary education), type and focus of performance task, type of rubrics used, measures of reliability and validity, measures of impact on students' learning, and students' and teachers' attitudes towards using rubrics as an assessment tool. We will first give an overview of the articles reviewed and analyze them according to the measurements used. Secondly, we will summarize the findings of how/if the use of rubrics effect students' learning and attitudes. The results will be presented in relation to each of the three research questions.

3. Results

3.1. Reliability of scoring

Most assessments have consequences for those being assessed (Black, 1998). Hence assessment has to be credible and trustworthy, and as such be made with disinterested judgment and grounded on some kind of evidence (Wiggins, 1998). Ideally, an assessment should be independent of who does the scoring and the results similar no matter when and where the assessment is carried out, but this is hardly obtainable. Whereas traditional testing, with for example multiple-choice questions, has been developed to meet more rigorous demands, complex performance assessment is being questioned on behalf of its credibility, focusing mainly on the reliability of the measurement. The more consistent the scores are over different raters and occasions, the more reliable the assessment is thought to be (Moskal & Leydens, 2000).

There are different ways in which variability in the assessment score can come up. It might be due to variations in the rater's (or raters') judgments, in students' performance (Black, 1998) or in the sampling of tasks (Shavelson, Gao, & Baxter, 1996). In this review we are mainly addressing the first source of variation, that is variation in judgment. It should be noted, however, that the other sources of variability might have a greater impact on reliability, and that task-sampling variability has been shown to be a serious threat to reliability in performance assessments (Shavelson et al., 1996).

Variations in raters' judgments can occur either across raters, known as inter-rater reliability, or in the consistency of one single rater, called intra-rater reliability. There are several factors that can influence the judgment of an assessor, and two raters might come to different conclusions about the same performance. Besides the more obvious reasons for disagreement, like differences in experience or lack of agreed-upon scoring routines, it has been reported that things like teachers' attitudes regarding students' ethnicity, as well as the content, may also influence the rating of students' work (Davidson, Howell, & Hoekema, 2000).

3.1.1. Intra-rater reliability

According to Brown, Bull, and Pendlebury (1997) the "major threat to reliability is the lack of consistency of an individual marker" (p. 235). Still only seven studies in this review have reported on intra-rater reliability.

Most of the studies investigating intra-rater reliability use Cronbach's alpha to estimate raters' consistency, and the majority[1] report on alpha values above .70, which, according to Brown, Glasswell, and Harland (2004), is generally

[1] Several studies in this review have computed more than one estimate but only report them on an aggregated level. This means that the precise number of estimates falling within a specified range, or exceeding a certain limit, cannot always be presented here.

Table 1
Overview of studies reporting on inter-rater reliability measurements

Method	Number of studies
Consensus agreements	
Percentage of total agreement	18
Percentage of adjacent agreement	14
Cohen's kappa	4
Other	7
Total[a]	27
Consistency estimates	
Pearson's correlations	4
Cronbach's alpha	8
Spearman's rho	6
Other	9
Total[a]	24
Measurement estimates	
Generalizability theory	15
Many-facets Rasch model	3
Other	1
Total[a]	19
Grand total[b]	46

[a] Some articles report on more than one method, for example both total and adjacent agreement. This category summarizes the total number of articles reporting on each inter-rater reliability measurement (e.g. consensus agreements) without counting any article twice.

[b] Some articles report on more than one inter-rater reliability measurement, for example both consistency and measurement estimates. This category summarizes the total number of articles reporting on inter-rater reliability without counting any article twice.

considered sufficient. As the same trend applies for studies using other estimates as well, this could indicate that intra-rater reliability might not in fact be a major concern when raters are supported by a rubric.

3.1.2. Inter-rater reliability

More than half of the articles in this review report on inter-rater reliability in some form, and many of these used percentage of agreement as a measurement. The frequent use of consensus agreements can probably be attributed to the fact that they are relatively easy to calculate, and that the method allows for the use of nominal data. Consistency estimates are measured mainly by means of different correlation coefficients, while the many-facets Rasch model and generalizability theory are the two main methods of measurement estimates. Below and in Table 1, the methods used, the range and the typical values of some indices are reported. It should be kept in mind that in a number of these studies, the intention has not been to generate typical values. Rather, manipulations have been made that might distort the values and thus the ranges of values should be interpreted with caution.

Mostly, percent of exact or adjacent agreement (within one score point) between raters is reported. The consensus estimates with percentage of exact agreement varied between 4 and 100% in this review, with the majority of estimates falling in the range of 55–75%. This means that many estimates failed to reach the criterion of 70% or greater, which is needed if exact agreement is to be considered reliable (Stemler, 2004). On the other hand, agreements within one score point exceeded 90% in most studies, which means a good level of consistency. It should be noted, however, that the consensus agreement of raters depends heavily on the number of levels in the rubric. With fewer levels, there will be a greater chance of agreement, and in some articles Cohen's kappa is used to estimate the degree to which consensus agreement ratings vary from the rate expected by chance. Kappa values between .40 and .75 represent fair agreement beyond chance (Stoddart, Abrams, Gasper, & Canaday, 2000), and the values reported vary from .20 to .63, with only a couple of values below .40.

When reporting on consistency estimates most researchers use some kind of correlation of raters' scoring, but in many articles it is not specified which correlation coefficient has been computed. Where specified, it is mostly Pearson's or Spearman's correlation, but in a few cases also Kendall's W. The range of correlations is .27–.98, with the majority between .55 and .75. In consistency estimates, values above .70 are deemed acceptable (Brown et al., 2004; Stemler, 2004), and, as a consequence, many fail to reach this criterion. Besides correlations between raters, the consistency is also reported with Cronbach's alpha. The alpha coefficients are in the range of .50–.92, with most values above .70.

Of the studies using measurement estimates to report on inter-rater reliability, generalizability theory has been used almost exclusively. A few have used the many-facets Rasch model and in one study ANOVA-based intraclass correlations has been used. Dependability and generalizability coefficients from generalizability theory ranged from .06 to .96 and from .15 to .98, respectively. Coefficient values exceeding .80 are often considered as acceptable (Brown et al., 2004; Marzano, 2002), but as most of the values reported were between .50 and .80, the majority of estimates do not reach this criterion.

3.1.3. Does the use of rubrics enhance the consistency of scoring?

Results from studies investigating intra-rater reliability indicate that rubrics seem to aid raters in achieving high internal consistency when scoring performance tasks. Studies focusing on this aspect of reliability are relatively few, however, as compared to those studying inter-rater reliability.

The majority of the results reported on rater consensus do not exceed 70% agreement. This is also true for articles presenting the consistency of raters as correlation coefficients, or as generalizability and dependability coefficients, where most of them are below .70 and .80, respectively. However, what is considered acceptable depends on whether the assessment is for high-stake or classroom purposes. A rubric that provides an easily interpretable picture of individual students' knowledge may lack technical quality for large-scale use. On the other hand, an assessment that provides highly reliable results for groups of students may fail to capture the performance of an individual student, and is then of no benefit to the classroom teacher (Gearhart, Herman, Novak, & Wolf, 1995). Also, while reliability could be seen as a prerequisite for validity in large-scale assessment, this is not necessarily true for classroom assessments. Decisions in the classroom, made on the basis of an assessment, can easily be changed if they appear to be wrong. Therefore, reliability is not of the same crucial importance as in large-scale assessments, where there is no turning back (Black, 1998). So, at least when the assessment is relatively low-stakes, lower levels of reliability can be considered acceptable. Consequently, most researchers in this review conclude that the inter-rater reliability of their rubrics is sufficient, even though the estimates are generally too low for traditional testing.

Of course, performance assessment *is* not traditional testing, and several findings in this review support the rather self-evident fact that when all students do the same task or test, and the scoring procedures are well defined; the reliability will most likely be high. But when students do different tasks, choose their own topics or produce unique items, then reliability could be expected to be relatively low (Brennan, 1996). An example is that extraordinary high correlations of rater scores are reported by Ramos, Schafer, & 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. However, there are several other factors influencing inter-rater reliability reported as well, which can be used to get a picture of how to make rubrics for performance assessments more reliable:

1. Benchmarks are most likely to increase agreement, but they should be chosen with care since the scoring depends heavily on the benchmarks chosen to define the rubric (Denner, Salzman, & Harris, 2002; Popp, Ryan, Thompson, & Behrens, 2003).
2. Analytical scoring is often preferable (Johnson, Penny, & Gordon, 2000; Johnson, Penny, & Gordon, 2001; Penny, Johnson, & Gordon, 2000a; Penny, Johnson, & Gordon, 2000b), but perhaps not so if the separate dimension scores are summarized in the end (Waltman, Kahn, & Koency, 1998).
3. Agreement is improved by training, but training will probably never totally eliminate differences (Stuhlmann, Daniel, Dellinger, Denny, & Powers, 1999; Weigle, 1999).
4. Topic-specific rubrics are likely to produce more generalizable and dependable scores than generic rubrics (DeRemer, 1998; Marzano, 2002).
5. Augmentation of the rating scale (for example that the raters can expand the number of levels using + or − signs) seems to improve certain aspects of inter-rater reliability, although not consensus agreements (Myford, Johnson, Wilkins, Persky, & Michaels, 1996; Penny et al., 2000a, 2000b). For high levels of consensus agreement, a two-level scale (for example competent–not competent performance) can be reliably scored with minimal training, whereas a four-level scale is more difficult to use (Williams & Rink, 2003).

6. Two raters are, under restrained conditions, enough to produce acceptable levels of inter-rater agreement (Baker, Abedi, Linn, & Niemi, 1995; Marzano, 2002).

In summary, as a rubric can be seen as a regulatory device for scoring, it seems safe to say that scoring with a rubric is probably more reliable than scoring without one. Furthermore, the reliability of an assessment can always, in theory, be raised to acceptable levels by providing tighter restrictions to the assessment format. Rubrics can aid this enhancement in the consistency of scoring by being analytic, topic-specific, and complemented with exemplars and/or rater training. The question then, is whether the changes brought about by these restrictions are acceptable, or if we lose the essence somewhere in the process of providing high levels of accuracy in scoring. Hence, even if important, reliability is not the only critical concept that has to be taken into account when designing performance assessments. The concept of validity must also be explored in relation to more authentic forms of assessment.

3.2. *Valid judgment of performance assessments*

Basically, validity in this context answers the question “Does the assessment measure what it was intended to measure?” The answer to this question, however, is not always as simple. There are two different ways of looking at validity issues. Either validity is seen as a property of the test, or as test score interpretations (Borsboom, Mellenbergh, & van Heerden, 2004; Brown et al., 1997). The first perspective is most widely used in natural sciences and psychological testing and no articles in this review were found using it. Instead, validity in educational research is often seen as to involve evaluative judgment, and is therefore not seen as a property of the test as such, but rather as an interpretation of the results (Borsboom et al., 2004; McMillan, 2004; Messick, 1996).

There are numerous aspects of validity investigated and reported in the literature on assessment. Most common are traditional criterion, content and construct validity. Messick (1996) argues for a more comprehensive theory of construct validity. He distinguishes six aspects of construct validity: content, generalizability, external, structural, substantive, and consequential. They may be discussed selectively, but none should be ignored.

The *content aspect* of Messick’s (1996) construct validity determines content relevance and representativeness of the knowledge and skills revealed by the assessment. Of the articles in this review, one-third reported on validity, and many of them used content validity in some way (see Table 2). Expert opinions were the number one route to get empirical evidence for this aspect of validity. The concern for content representativeness in assessment is due to the need for results to be generalizable to the construct domain, and not be limited to only the sample of assessed tasks. Messick (1996) distinguishes two aspects of validity in this regard, where the *generalizability aspect* refers to the extent to which score interpretations generalize across groups, occasions, tasks, etc., while the *external aspect* examines the relationship of the assessment score to other measures relevant to the construct being assessed. The boundary between the generalizability aspect and the external aspect seems to be somewhat unclear in some instances, as when Baker (1994) makes comparisons of her assessment with other tests, grades and evaluations. However, making comparisons of students’ scores across grade levels, like Gearhart et al. (1995), could probably be referred to as addressing the generalizability aspect of construct validity.

Table 2
Overview of studies reporting on rubric validity

Aspect of validity[a]	Number of studies
Content	10
Generalizability	3
External	15
Structural	7
Substantive	1
Consequential	2
Total[b]	25

[a] Adapted from Messick (1996).

[b] Some articles report on more than one aspect and this category summarizes the total number of articles reporting on each aspect without counting any article twice.

Reporting on external aspects of validity, several articles use correlations with other measures or instruments, such as an established rubric (Gearhart et al., 1995), post-course evaluations (Roblyer & Wiencke, 2003), national survey items (Stoering & Lu, 2002) or tests of prior knowledge (Waltman et al., 1998). Most report on modest correlations from .4 to .6.

Another focus of external validity is the relevance and utility of the rubric for its intended purpose. For instance, Flowers and Hancock (2003) report that their interview protocol and scoring rubric for evaluating teacher performance has been adopted by over 85% of the public schools of North Carolina.

In the beginning of this article, Perlman (2003) was cited, saying that performance assessment consist of two parts: "a task and a set of scoring criteria or a scoring rubric" (p. 497). According to Messick (1996), not only does the task have to be consistent with the theory of the construct in question, but the scoring structure (like criteria and rubric) must also follow rationally from the domain structure. This is called the *structural aspect* of construct validity and has been addressed in some studies by means of factor analysis (Baker, 1994; Baker et al., 1995) and by raters evaluating the alignment of guidelines, standards and the rubric (Denner et al., 2002).

As mentioned above, the content aspect was a frequently investigated aspect of validity, and empirical evidence was mainly collected through expert opinions. Domain coverage is not only about traditional content, however, but also about thinking processes used during the assessment. The *substantive aspect* includes theoretical rationales for, and empirical evidence of, consistency in responses that reflect the thinking processes used by experts in the field. Also, attention has to be paid to the level of these cognitive processes in the assessment (Van de Watering & van der Rijt, 2006). In the studies reviewed, most rubrics focus on products, like essays, work samples or laboratory reports, rather than processes. A noteworthy exception is the relatively large amount of studies investigating student teachers. As an example, Osana and Seymour (2004) designed a rubric according to empirically validated theory in argumentation and statistical reasoning. Therefore, the rubric could serve not only as an assessment tool, but also as "a theoretical model of good thinking" (p. 495).

The last aspect of validity, the *consequential aspect*, includes evidence of implications of score interpretation, both intended and unintended as well as short- and long-term consequences (Messick, 1996). Only two articles in this review reports explicitly on *consequential aspects* of validity and one of them is a study by Gearhart et al. (1995), where they try to validate a new rubric for narrative writing. The researchers are guided in this validation process by the work of Messick and use an established rubric for comparison. Under the headings of "value implications" and "consequential validity" they examine evidence from raters' reflections of score usefulness in informing writing instruction as well as the stability and meaning of decisions of mastery based on different cut points. Value implications, as suggested by raters' reflections, indicate that the new rubric has more instructional potential than the comparison rubric. When considering social consequences of decisions about mastery/non-mastery on scores derived from the rubric, the authors discuss the possibility that some individuals might be judged differently based on the two rubrics.

3.2.1. Can rubrics facilitate valid judgment of performance assessments?

Most reports claim to have some support for the validity of the rubric used. Several rubrics have been validated for content validity by experts, and also the scores produced have been checked for correlation to other measures, both internal and external. Researchers have performed factor analysis to reveal the underlying structure or investigated the alignment of guidelines, standards and rubrics. Only one study, Gearhart et al. (1995), has used a more comprehensive framework for the validation process.

It is still 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 thinking processes, 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, in the sense that students are disadvantaged in their opportunity to show what they have learned.

On the issue of reliability it was concluded that, since a rubric is a regulatory device, scoring with a rubric is probably more reliable than scoring without. Could it, in the same sense, be concluded that scoring with a rubric is probably more valid than scoring without? The answer in this case would have to be "no". Just by providing a rubric there is no evidence for content representativeness, fidelity of scoring structure to the construct domain or generalizability. Nor does it give any convergent or discriminant evidence to other measures. There is, however, one certain aspect of validity that might benefit from the use of rubrics. If rubrics in some way affect instruction, so that there are positive educational consequences from using them, then this would influence the aspect of consequential validity.

Table 3
Overview of studies reporting on promotion of student learning and/or the quality of teaching

Data	Number of studies
Student improvement	10
Teachers' perceptions	9
Students' perceptions	8
Student use of criteria and self-assessment	8
Other	2
Total[a]	25

[a] Some articles report on more than one category of data and here the total number of articles reporting on each category is summarized, without counting any article twice.

3.3. *Promotion of student learning and/or the quality of teaching*

As is widely recognized, assessment has a strong impact on the focus and attention of most students (Dochy et al., 2006). This, up till now mostly negative influence, has been acknowledged by educational institutions, and has led to the demand for more authentic, complex, and motivating forms of assessment. The performance movement and the investment in rubrics are part of this. There is a strong conviction that the use of performance assessment in combination with rubrics will change students' efforts and learning in a positive way. In line with this assumption, all of the reviewed articles argue that the use of rubrics has shown to be beneficial for students' learning, motivation and study situation at large.

Performance assessments are by definition open-ended, and the plethora of outcomes is not easily predicted, let alone measured with high accuracy. In this sense, the persons in best position to evaluate if rubrics promote learning and/or improve instruction are the students and teachers actually using them. Hence both students' and teachers' perceptions of educational consequences are presented alongside more solid research results in this review.

Of the 75 articles reviewed, one-third report on some kind of educational consequences of rubric usage, the majority being concerned with student improvement or/and perceptions of using rubrics by either teachers, students or both. Eight articles investigated the effect of rubrics on self- and peer assessment, and a few articles also reported on the effect of rubrics on off-task behaviour, evaluation time and students' understanding of criteria (see Table 3).

3.3.1. *Self- and peer assessment*

The research literature on self- and peer assessment is substantive. It is claimed that it is advantageous for students' learning to be involved in giving and receiving feedback (Dochy, Segers, & Sluijsmans, 1999; Topping, 2003). The meta-analyses of Falchikov and Boud (1989) and of Falchikov and Goldfinch (2000) provide a comprehensive list of research on self- and peer assessment. There are, however, few scientific studies reporting on effects of self- and peer assessment using rubrics. Schirmer, Bailey, and Fitzgerald (1999) report on a year long experiment with assessment rubrics as a teaching strategy with deaf children in the fifth and seven grades, where significant improvement in the quality of students' compositions were made. The evaluation of student improvement was done using both quantitative and qualitative measures. The quantitative analysis indicated that use of the rubric as a teaching strategy, significantly improved writing according to topic, content, story development and organization.

Research on self- and peer assessment at large indicates on the one hand that students can be very accurate in grading their own work (Dochy et al., 1999), whereas on the other hand self-assessment tend to result in higher grades than teacher assessment (Topping, 2003). Taken together, it seems as if assessment of one's own performance is more difficult than assessing a peer's performance (Lindblom-Ylänne, Pihlajamäki, & Kotkas, 2006). A central question that has to be further evaluated, is if the use of rubrics might enhance the accuracy of self- and peer assessment.

The meta-analyses mentioned above, no ted low technical quality regarding the quantitative research reviewed. A few recent articles have investigated the variation in students' responses in relation to a scoring rubric using quantitative measures (e.g. Cho, Schunn, & Wilson, 2006; Hafner & Hafner, 2003; Sadler & Good, 2006).

The study by Sadler and Good (2006) puts the presumed benefits of peer-grading to the test. The researchers compared teacher-assigned grades to grades awarded either by students to themselves or by their peers. Students in four middle school classrooms were trained to grade with the help of a scoring rubric. A very high correlation between students and their teacher was obtained (.91–.94). An important finding was that the students who scored their own tests using the rubric, improved dramatically. The authors conclude that both self- and peer-grading may be used to save teachers' time on grading, and also that self-grading appears to result in increased student learning, whereas peer-grading does not.

Hafner and Hafner (2003) used assessments of oral presentations to estimate the reliability of a rubric for self- and peer assessment purposes. When supported by a rubric, the students showed much agreement in their ranking of the presentations.

Cho et al. (2006) argue that peer reviewing of writing may be a way to create more writing opportunities in college and university settings, but observes that the validity and reliability of peer-generated grades are a major concern. Their analysis suggests that the aggregated ratings of at least four peers are both highly reliable and as valid as instructor ratings.

Although few, these studies indicate that rubrics might be valuable in supporting student self- and peer assessment. Some studies also show that students actually internalize the criteria, making them their own, and use them while self-assessing (Andrade, 1999b; Piscitello, 2001).

3.3.2. Student improvement and users perceptions

It is not possible to draw any conclusions about student improvement related to the use of rubrics from this material. This is mainly due to the fact that the results are not pointing in any particular direction. In the studies reporting on student improvement of some kind, only two show an overall improvement (Brown et al., 2004; Mullen, 2003) while others have positive effects only in some areas (Green & Bowser, 2006; Sadler & Good, 2006; Schafer, Swanson, Bené, & Newberry, 2001; Schamber & Mahoney, 2006; Schirmer et al., 1999), or only in combination with other interventions (Toth, Suthers, & Lesgold, 2002), and in one study some negative effects (Andrade, 1999a). The perceptions of the users as to the benefits of using rubrics may therefore be seen as more interesting. A major theme in the comments from both teachers and students, is the perception of clarified expectations or, in the wording by Frederiksen and Collins (1989), *transparency*. Rubrics indicate what is important and thereby give clarity and explicitness to the assessment, and this is deemed positive by students and teachers alike (Bissell & Lemons, 2006; Morrell & Ackley, 1999; Schamber & Mahoney, 2006; Shaw, 2004).

Besides transparency, other benefits of rubrics as perceived by the teachers are the encouragement of *reflective practice* (Beeth et al., 2001; Luft, 1998), and that rubrics can give teachers more insights to the *effectiveness* of their instructional practices (Waltman et al., 1998). Also, the concrete nature of rubric criteria provides information for *feedback* as well as makes *self-assessment* easier (Schamber & Mahoney, 2006; Smith & Hanna, 1998).

A couple of studies report on activity and off-task behaviour, where students seem more involved in the task at hand (Piscitello, 2001; Toth et al., 2002). One possible interpretation of this is that rubrics help, via transparency, to make assignments and assessment meaningful to the students. They know why they are doing what they are doing.

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

To conclude, it seems like the use of rubrics have 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.

4. Discussion

The distribution of the reviewed articles, with a majority tackling questions of reliability and validity, and a few working with the effects of using rubrics on students' learning and the quality of the teaching–learning situation, mirrors in a way the widespread interest in assessing performance in a credible way. The relative lack of research studies on the effects of learning and teaching does not, however, mirror the great expectations and positive narratives of the effect of rubrics on the quality of performance assessment. Even if research articles have been presented on the topic for a decade, the research may still be described as rudimentary.

The studies reporting on rubric reliability, reviewed in this article, generally present low reliability coefficients as compared to traditional psychometric requirements, indicating that the use of rubrics might not in itself be enough to produce sufficient reliability for summative assessments. Performance assessment is, however, open-ended and as such prone to produce lower reliability. Still, reliability can be improved by adding restrictions to the assessment.

Benchmarks can be used, raters can be trained, different scoring methods can be applied, etc. Since rubrics are a way of restricting the scoring of performance assessments, the use of rubrics should in most cases improve reliability.

Rather, the question is: If severe restrictions are made, due to calls for high reliability, do we still measure the full scope of what was intended to measure? In this view, reliability is not the "bottleneck" for quality performance assessments, but rather validity seems to be of more critical importance.

However, validity issues are not always straight forward. The validity concept has traditionally been fragmented into different forms of validity, for example criterion and content validity. In a more contemporary view of validity, the term construct validity refers to a unifying concept incorporating different aspects of validity. All these aspects are seen as interrelated, and all should be addressed when validating assessments, in order to get a more complete picture of the validity.

This holistic, broader approach has not been used by most articles. Typically, one or two aspects of validity have been addressed while the others are left unmentioned. Thus, as an example, it is not known whether an assessment deemed valid for correlating with external measurements actually requires the higher order thinking that was intended.

The evidence for student improvement due to the usage of rubrics is still scarce if we restrict ourselves to rigorous quantitative studies. Performance assessments target knowledge and skills which are often difficult to evaluate with the traditional pre- and post-tests of educational research. The studies reviewed do, however, present positive results. In line with the assumptions from research on self- and peer assessment, they indicate that learning is promoted by the meta-cognitive processes involved in this type of assessments, which could be further aided by the use of rubrics. A few of these studies are long term and involve many students. As the type of content involved, as well as the performance tasks assessed, seem to be important factors influencing the results, generalization of the data is still not recommended. Evaluations of teachers' and students' experiences and attitudes are on the contrary almost univocal, and positive. Those actors are perhaps in best positions to evaluate the benefits or detriments of using rubrics.

The reviewed research on teachers' and students' perceptions of using rubrics, shows that a major benefit of rubrics is that of bringing transparency to the assessment, which makes expectations explicit. The question has been raised as to whether the transparency provided by rubrics could actually stifle creativity (Mabry, 1999; Messick, 1996). To avoid this, Wiggins (1998) emphasizes that rubrics should not restrict the format or the method. By using various examples or "anchors" it is also possible to show that there are many ways to approach the same task.

Student understanding of criteria, feedback, possibilities of self- and peer assessment are other positive experiences reported. So even if it is not strictly demonstrated that students do learn better, the students themselves perceive that they do. Knowing that learning is influenced by factors such as motivation (Birenbaum et al., 2006), transparency of assessments can be seen as a great contributor to learning.

The question of reliability versus validity is actualized when the effects on student learning is studied. In addition to transparency there are a row of other possible benefits of the use of rubrics for performance assessment. The question is addressed in a review on requirements for competence assessments (Baartman, Bastiaens, Kirschner, & Van der Vleuten, submitted for publication). The researchers argue that transparency is related to both the structural as well as the consequential aspect of validity. Alignment, authenticity and other similar concepts are in the same way demonstrated to be associated with Messick's (1996) framework. This is in line with the argument of some researchers that novel forms of assessment cannot be evaluated only on the basis of psychometric criteria. Instead a new, or at least a widened, set of criteria is needed, reflecting the qualities sought in new modes of assessment (Gielen, Dochy, & Dierick, 2003). Baartman et al. (submitted for publication) offer a way to resolve the issue by meeting the demands of both psychometricians and the emerging "assessment culture". This is done by suggesting that multimodal assessment programs should be developed for high-stakes competence assessments, where different methods of assessment are combined. In this way, not each individual assessment has to meet all criteria, but by using a combination of methods the program as a whole can meet the quality criteria of both cultures.

Baartman, Bastiaens, Kirschner, & Van der Vleuten (2006) presents a framework containing twelve quality criteria for competence assessment programs. The authors put forth what is called "The Wheel of Competency Assessment", where the quality criteria are displayed in concentric circles. The hub is occupied by "Fitness for purpose"—the foundation of assessment development, surrounded by "Comparability", "Reproducibility of decisions", "Transparency" and "Acceptability". These basic criteria are seen as prerequisites for the outer layer, consisting of "Fairness", "Fitness for self-assessment", "Meaningfulness", "Cognitive complexity" and "Authenticity". The wheel, with its ten criteria, is framed within an educational context, represented by two criteria: "Educational consequences" and "Costs & Efficiency". With such a framework there is no need to take the detour, via the traditional psychometric criteria reliability and validity, in order to estimate the quality of performance assessments.

5. Conclusions

This paper aimed to review empirical research and illuminate the questions of how the use of rubrics can (1) enhance the reliability of scoring, (2) facilitate valid judgment of performance assessments, and (3) give positive educational consequences, such as promoting learning and/or improve instruction.

A first conclusion is that the reliable scoring of performance assessments can be enhanced by the use of rubrics. In relation to reliability issues, rubrics should be analytic, topic-specific, and complemented with exemplars and/or rater training. Since performance assessments are more or less open ended per definition, it is not always possible to restrict the assessment format to achieve high levels of reliability without sacrificing the validity.

Another conclusion is that rubrics do not facilitate valid judgment of performance assessments per se. Valid assessment could be facilitated by using a more comprehensive framework of validity when validating the rubric, instead of focusing on only one or two aspects of validity. In relation to learning and instruction, consequential validity is an aspect of validity that might need further attention.

Furthermore, it has been concluded that rubrics seem to have the potential of promoting learning and/or improve instruction. The main reason for this potential lies in the fact that rubrics make expectations and criteria explicit, which also facilitates feedback and self-assessment. It is thus argued that assessment quality criteria should emphasize dimensions like transparency and fitness for self-assessment to a greater extent than is done through the traditional reliability and validity criteria. This could be achieved through a framework of quality criteria that acknowledges the importance of trustworthiness in assessment as well as supports a more comprehensive view on validity issues (including educational consequences).

References

Arter, J., & McTighe, J. (2001). *Scoring rubrics in the classroom*. Thousand Oaks: Corwin Press Inc..

Baartman, L. K. J., Bastiaens, T. J., Kirschner, P. A., & Van der Vleuten, C. P. M. (submitted for publication). Assessment in competence-based education: How can assessment quality be evaluated? *Educational Research Review*.

Baartman, L. K. J., Bastiaens, T. J., Kirschner, P. A., & Van der Vleuten, C. P. M. (2006). The wheel of competency assessment: Presenting quality criteria for competency assessment programmes. *Studies in Educational Evaluation*, *32*, 153–170.

Birenbaum, M., Breuer, K., Cascallar, E., Dochy, F., Dori, Y., Ridgway, J., et al. (2006). A learning integrated assessment system. *Educational Research Review*, *1*, 61–67.

Black, P. (1998). *Testing: Friend or foe?* London: Falmer Press.

Borsboom, D., Mellenbergh, G. J., & van Heerden, J. (2004). The concept of validity. *Psychological Review*, *111*, 1061–1071.

Brennan, R. (1996). Generalizability of performance assessments. In G. Phillips (Ed.), *Technical issues in large-scale performance assessment* (pp. 19–58). Washington, DC: National Center for Education Statistics.

Brown, G., Bull, J., & Pendlebury, M. (1997). *Assessing student learning in higher education*. London: Routledge.

Busching, B. (1998). Grading inquiry projects. *New Directions for Teaching and Learning*, 89–96.

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

Davidson, M., Howell, K. W., & Hoekema, P. (2000). Effects of ethnicity and violent content on rubric scores in writing samples. *Journal of Educational Research*, *93*, 367–373.

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.

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

Falchikov, N., & Goldfinch, J. (2000). Student peer assessment in higher education: A meta-analysis comparing peer and teacher marks. *Review of Educational Research*, *70*, 287–322.

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*. Dordrecht: Kluwer Academic Publishers.

Mabry, L. (1999). Writing to the rubric: Lingering effects of traditional standardized testing on direct writing assessment. *Phi Delta Kappan*, *80*, 673–679.

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

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

Morrison, G. R., & Ross, S. M. (1998). Evaluating technology-based processes and products. *New Directions for Teaching and Learning*, *74*, 69–77.

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

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

Perlman, C.C. (2003). *Performance assessment: Designing appropriate performance tasks and scoring rubrics*. North Carolina, USA.

Shavelson, R. J., Gao, X., & Baxter, G. (1996). On the content validity of performance assessments: Centrality of domain-specifications. In M. Birenbaum & F. Dochy (Eds.), *Alternatives in assessment of achievements, learning processes and prior knowledge*. Boston: Kluwer Academic Publishers.

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

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*. Dordrecht: Kluwer Academic Publishers.

Van de Watering, G., & van der Rijt, J. (2006). Teachers' and students' perceptions of assessments: A review and a study into the ability and accuracy of estimating the difficulty levels of assessment items. *Educational Research Review*, *1*, 133–147.

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

References to papers in the review

Andrade, H. G. (1999a). The role of instructional rubrics and self-assessment in learning to write: A smorgasbord of findings. In *Paper Presented at the Annual Meeting of the American Educational Research Association*.

Andrade, H. G. (1999b). Student self-assessment: At the intersection of metacognition and authentic assessment. In *Paper Presented at the Annual Meeting of the American Educational Research Association*.

Baker, E. L. (1994). Learning-based assessments of history understanding. *Educational Psychologist*, *29*, 97–106.

Baker, E. L., Abedi, J., Linn, R. L., & Niemi, D. (1995). Dimensionality and generalizability of domain-independent performance assessments. *Journal of Educational Research*, *89*, 197–205.

Beeth, M. E., Cross, L., Pearl, C., Pirro, J., Yagnesak, K., & Kennedy, J. (2001). A continuum for assessing science process knowledge in grades K-6. *Electronic Journal of Science Education*, 5.

Bissell, A. N., & Lemons, P. R. (2006). A new method for assessing critical thinking in the classroom. *BioScience*, *56*, 66–72.

Brown, G. T. L., Glasswell, K., & Harland, D. (2004). Accuracy in the scoring of writing: Studies of reliability and validity using a new zealand writing assessment system. *Assessing Writing*, *9*, 105–121.

Chi, E. (2001). Comparing holistic and analytic scoring for performance assessment with many-facet rasch model. *Journal of Applied Measurement*, *2*, 379–388.

Cho, K., Schunn, C. D., & Wilson, R. W. (2006). Validity and reliability of scaffolded peer assessment of writing from instructor and student perspectives. *Journal of Educational Psychology*, *98*, 891–901.

Denner, P. R., Salzman, S. A., & Harris, L. B. (2002). Teacher work sample assessment: An accountability method that moves beyond teacher testing to the impact of teacher performance on student learning. In *Paper presented at the annual meeting of the American Association of Colleges for Teacher Education*.

DeRemer, M. L. (1998). Writing assessment: Raters' elaboration of the rating task. *Assessing Writing*, *5*, 7–29.

Duke, B. 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. USA: University of Oklahoma.

Flowers, C. P., & Hancock, D. R. (2003). An interview protocol and scoring rubric for evaluating teacher performance. *Assessment in Education: Principles, Policy and Practice*, *10*, 161–168.

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

Green, R., & Bowser, M. (2006). Observations from the field: Sharing a literature review rubric. *Journal of Library Administration*, *45*, 185–202.

Hafner, J. C., & Hafner, P. M. (2003). Quantitative analysis of the rubric as an assessment tool: An empirical study of student peer-group rating. *International journal of science education*, *25*, 1509–1528.

Johnson, R. L., Penny, J., & Gordon, B. (2000). The relation between score resolution methods and interrater reliability: An empirical study of an analytic scoring rubric. *Applied Measurement in Education*, *13*, 121–138.

Johnson, R. L., Penny, J., & Gordon, B. (2001). Score resolution and the interrater reliabilityof holistic scores in rating essays. *Written Communication*, *18*, 229–249.

Lindblom-Ylänne, S., Pihlajamäki, H., & Kotkas, T. (2006). Self-, peer- and teacher-assessment of student essays. *Active Learning in Higher Education*, *7*, 51–62.

Luft, J. (1998). Rubrics: Design and use in science teacher education. In *Paper Presented at the Annual Meeting of the Association for the Education of Teachers in Science*.

Lunsford, B. E. (2002). *Inquiry and inscription as keys to authentic science instruction and assessment for preservice secondary science teachers*. Unpublished doctoral dissertation. USA: University of Tennessee.

Marzano, R. J. (2002). A comparison of selected methods of scoring classroom assessments. *Applied Measurement in Education*, *15*, 249–267.

McMartin, F., McKenna, A., & Youssefi, K. (2000). Scenario assignments as assessment tools for undergraduate engineering education. *IEEE Transactions on Education*, *43*, 111–120.

Morrell, P. D., & Ackley, B. C. (1999). Practicing what we teach: Assessing pre-service teachers' performance using scoring guides. In *Paper presented at the annual meeting of the American Educational Research Association*.

Mullen, Y. K. (2003). *Student improvement in middle school science.* Unpublished master thesis. USA: University of Wisconsin.

Myford, C. M., Johnson, E., Wilkins, R., Persky, H., & Michaels, M. (1996). Constructing scoring rubrics: Using "facets" to study design features of descriptive rating scales. In *Paper presented at the annual meeting of the American Educational Research Association*.

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

Paratore, J. R. (1995). Assessing literacy: Establishing common standards in portfolio assessment. *Topics in Language Disorders*, *16*, 67–83.

Penny, J., Johnson, R. L., & Gordon, B. (2000a). The effect of rating augmentation on inter-rater reliability: An empirical study of a holistic rubric. *Assessing Writing*, *7*, 143–164.

Penny, J., Johnson, R. L., & Gordon, B. (2000b). Using rating augmentation to expand the scale of an analytic rubric. *Journal of Experimental Education*, *68*, 269–287.

Piscitello, M. E. (2001). *Using rubrics for assessment and evaluation in art.* Unpublished master thesis. USA: Saint Xavier University.

Popp, S. E. O., Ryan, J. M., Thompson, M. S., & Behrens, J. T. (2003). Operationalizing the rubric: The effect of benchmark selection on the assessed quality of writing. In *Paper Presented at Annual Meeting of the American Educational Research Association*.

Ramos, K. D., Schafer, S., & Tracz, S. M. (2003). Validation of the fresno test of competence in evidence based medicine. *British Medical Journal*, *326*, 319–321.

Roblyer, M. D., & Wiencke, W. R. (2003). Design and use of a rubric to assess and encourage interactive qualities in distance courses. *American Journal of Distance Education*, *17*, 77–99.

Sadler, P. M., & Good, E. (2006). The impact of self- and peer-grading on student learning. *Educational Assessment*, *11*, 1–31.

Schafer, W. D., Swanson, G., Bené, N., & Newberry, G. (2001). Effects of teacher knowledge of rubrics on student achievement in four content areas. *Applied Measurement in Education*, *14*, 151–170.

Schamber, J. F., & Mahoney, S. 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.

Schirmer, B. R., Bailey, J., & Fitzgerald, S. M. (1999). Using a writing assessment rubric for writing development of children who are deaf. *Exceptional Children*, *65*, 383–397.

Shaw, J. (2004). Demystifying the evaluation process for parents: Rubrics for marking student research projects. *Teacher Librarian*, *32*, 16–19.

Smith, J., & Hanna, M. A. (1998). Using rubrics for documentation of clinical work supervision. *Counselor Education and Supervision*, *37*, 269–278.

Stoddart, T., Abrams, R., Gasper, E., & Canaday, D. (2000). Concept maps as assessment in science inquiry learning—A report of methodology. *International Journal of Science Education*, *22*, 1221–1246.

Stoering, J. M., & Lu, L. (2002). Combining the national survey of student engagement with student portfolio assessment. In *Paper Presented at Annual Meeting of the Association for Institutional Research*.

Stuhlmann, J., Daniel, C., Dellinger, A., Denny, R. K., & Powers, T. (1999). A generalizability study of the effects of training on teachers' abilities to rate children's writing using a rubric. *Journal of Reading Psychology*, *20*, 107–127.

Toth, E. E., Suthers, D. D., & Lesgold, A. M. (2002). Mapping to know: The effects of representational guidance and reflective assessment on scientific inquiry. *Science Education*, *86*, 264–286.

Waltman, K., Kahn, A., & Koency, G. (1998). *Alternative approaches to scoring: The effects of using different scoring methods on the validity of scores from a performance assessment*. *CSE Technical Report 488*. Los Angeles.

Weigle, S. C. (1999). Investigating rater/prompt interactions in writing assessment: Quantitative and qualitative approaches. *Assessing Writing*, *6*, 145–178.

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

References to papers in the review

Abedi, J., & Baker, E. L. (1995). A latent-variable modeling approach to assessing interrater reliability, topic generalizability, and validity of a content assessment scoring rubric. *Educational and Psychological Measurement*, *55*, 701–715.

Aschbacher, P. R. (1999). *Developing indicators of classroom practice to monitor and support school reform*. *CSE Technical Report 513*. Los Angeles.

Borko, H., & Stecher, B. (2006). *Using classroom artifacts to measure instructional practice in middle school science*: *A two-state field test*. *CSE Technical Report 690*. Los Angeles.

Boston, M., & Wolf, M. K. (2006). *Assessing academic rigor in mathematics instruction*: *The development of the instructional quality assessment toolkit*. *CSE Technical Report 672*. Los Angeles.

Brookhart, S. M. (2005). The quality of local district assessments used in Nebraska's school-based teacher-led assessment and reporting system (STARS). *Educational Measurement: Issues and Practice*, *24*, 14–21.

Choinski, E., Mark, A. E., & Murphey, M. (2003). Assessment with rubrics: An efficient and objective means of assessing student outcomes in an information resources class. *Portal: Libraries & the Academy*, *3*, 563–576.

Clare, L., Valdes, R., Pascal, J., & Steinberg, J. R. (2001). *Teachers' assignments as indicators of instructional quality in elementary schools*. *CSE Technical Report 545*. Los Angeles.

Dunbar, N. E., Brooks, C. F., & Kubicka-Miller, T. (2006). Oral communication skills in higher education: Using a performance-based evaluation rubric to assess communication skills. *Innovative Higher Education*, *31*, 115–128.

Flowers, C. (2006). Confirmatory factor analysis of scores on the clinical experience rubric. *Educational and Psychological Measurement*, *66*, 478–488.

Goldberg, G. L., Roswell, B. S., & Michaels, H. (1998). A question of choice: The implications of assessing expressive writing in multiple genres. *Assessing Writing*, *5*, 39–70.

Harrison, J. A., McAfee, H., & Caldwell, A. (2002). Examining, developing, and validating the interview for admission into the teacher education program. In *Paper Presented at the Annual Meeting of the Southeastern Region Association for Teacher Educators*.

Hickey, D. T., DeCuir, J., Hand, B., Kyser, B., Laprocina, S., & Mordica, J. (2002). *Technology-supported formative and summative assessment of collaborative scientific inquiry*. Learning & Performance Support Laboratory, University of Georgia.

Johnson, R. L., Fisher, S., Willeke, M. J., & McDaniel, F. (2003). Portfolio assessment in a collaborative program evaluation: The reliability and validity of a family literacy portfolio. *Evaluation and Program Planning*, *26*, 367–377.

Keiser, J. C., Lawrenz, F., & Appleton, J. J. (2004). Technical education curriculum assessment. *Journal of Vocational Education Research*, *29*, 181–194.

Koul, R., Clariana, R. B., & Salehi, R. (2005). Comparing several human and computer-based methods for scoring concept maps and essays. *Journal of Educational Computing Research*, *32*, 227–239.

Laveault, D., & Miles, C. (2002). The study of individual differences in the utility and validity of rubrics in the learning of writing ability. In *Paper Presented at the Annual Meeting of the American Educational Research Association*.

Matsumura, L. C., Slater, S. C., Wolf, M. K., Crosson, A., Levison, A., Peterson, M., Resnick, L., & Junker, B. (2006). *Using the instructional quality assessment toolkit to investigate the quality of reading comprehension assignments and student work. CSE Technical Report 669*. Los Angeles.

Mott, M. S., Etsler, C., & Drumgold, D. (2003). Applying an analytic writing rubric to children's hypermedia "narratives". *Early Childhood Research & Practice: An Internet Journal on the Development, Care, and Education of Young Children*, *5*.

Olson, L., Schieve, A. D., Ruit, K. G., & Vari, R. C. (2003). Measuring inter-rater reliability of the sequenced performance inventory and reflective assessment of learning (SPIRAL). *Academic Medicine*, *78*, 844–850.

Pindiprolu, S. S., Lignugaris/Kraft, B., Rule, S., Peterson, S., & Slocum, T. (2005). Scoring rubrics for assessing students' performance on functional behavior assessment cases. *Teacher Education and Special Education*, *28*, 79–91.

Pomplun, M., Capps, L., & Sundbye, N. (1998). Criteria teachers use to score performance items. *Educational Assessment*, *5*, 95–110.

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

Scherbert, T. G. (1998). *Collaborative consultation pre-referral interventions at the elementary level to assist at-risk students with reading and language arts difficulties*. Unpublished doctoral dissertation. USA: Nova Southeastern University.

Simon, M., & Forgette-Giroux, R. (2001). A rubric for scoring postsecondary academic skills. *Practical Assessment, Research & Evaluation*, *7*.

Ward, J. R., & McCotter, S. S. (2004). Reflection as a visible outcome for preservice teachers. *Teaching and Teacher Education*, *20*, 243–257.

Watkins, T. J. (1996). *Validity and internal consistency of two district-developed assessments of Title I students.* Unpublished doctoral dissertation. USA: University of Illinois at Urbana-Champaign.