Educational Review, 2015
Vol. 67, No. 3, 343–368, http://dx.doi.org/10.1080/00131911.2014.929565

Routledge
Taylor & Francis Group

The quality and effectiveness of descriptive rubrics

Susan M. Brookhart[a]* and Fei Chen[b]

[a]*Duquesne University, Pittsburgh, PA, USA;* [b]*State University of New York, University at Albany, Albany, NY, USA*

This review synthesizes the findings of studies of the use of rubrics in education settings published from 2005 to 2013. The review included studies only if the rubrics involved met the definition of having coherent sets of criteria and performance level descriptions for those criteria. Compared to the results of a previous review by Jonsson and Svingby (*Educational Research Review* 2(2): 130–144, 2007), the frequency, scope, and rigor of studies of rubrics have increased in recent years. Rubrics yield information of sufficient quality if certain conditions are met, most notably having clear and focused criteria. Evidence regarding the effects of rubrics on performance is positive overall. Evidence of the effects of rubrics on self-regulation of learning is mixed, though positive associations between rubric use and motivation to learn were identified in some studies.

Keywords: rubrics; validity; reliability; student learning; motivation to learn

Introduction

A rubric is a coherent set of criteria for students' work that includes descriptions of levels of performance quality on the criteria (Brookhart 2013b). Rubrics by this definition differ from rating scales, which have criteria but no performance level descriptions, although in common parlance these are often called "rubrics". Several important purposes claimed for rubrics, that they facilitate student self-assessment, facilitate teacher and peer feedback, and help students envision what to do to improve their work, only make sense if both criteria and performance level descriptions are present (Andrade 2000; Jonsson and Svingby 2007). The purpose of this article is to document what is known about (1) the quality of assessment information from rubrics and (2) the effects of rubric use on student learning and motivation to learn, using the method of a literature review.

Rubric use in educational settings

Rubrics enjoy wide use in primary, secondary, and post-secondary education. They arose as part of a response to research in the 1980s suggesting that students were better at repeating facts and concepts than applying them, and the consequent interest in performance assessment (Lane and Tierney 2008) and standards-based reform (Brookhart 2013a). Performance assessment can reflect students' abilities to solve real-world problems, analyze and synthesize information, and apply their knowledge and skills only if they have carefully designed scoring procedures with clear criteria.

*Corresponding author. Email: susanbrookhart@bresnan.net

Rubrics with criteria and performance level descriptions are deemed an effective vehicle for organizing and communicating criteria and performance expectations and for use in scoring and/or providing feedback on student work (Lane and Tierney 2008).

Beginning in the 1990s, two other developments have bolstered the use of rubrics: increasing emphasis on formative assessment in education at all levels (Andrade and Cizek 2010) and increasing emphasis on assessment and evaluation in the accreditation process in post-secondary education (Gerretson and Golson 2005; Kerby and Romine 2010). Formative assessment and accreditation may seem strange bedfellows, since accreditation relies heavily on summative assessment data. However, they both rely on clear statements of criteria for the quality of student work, and therefore they both provide an impetus for rubric use despite their differences in assessment purpose. These diverse sources of energy are fueling a growth in rubric use at present and are the justification for reviewing what is known about the quality and effectiveness of rubric use in educational settings.

Three literature reviews on the topic of rubrics (Jonsson and Svingby 2007; Panadero and Jonsson 2013; Reddy and Andrade 2010) have been published recently. Jonsson and Svingby (2007) did a comprehensive literature review; the present review updates their findings. The present review is broader than the other two reviews. Panadero and Jonsson (2013) specifically addressed the use of rubrics in formative assessment; the present review includes studies of both formative and summative use of rubrics. Reddy and Andrade (2010) addressed the use of rubrics in post-secondary education; the present review includes studies of rubric use in all educational settings.

Method

An electronic search of five databases was conducted, seeking peer-reviewed journal articles published from 2005 through October 2013. The five databases were Academic Search Complete (ASC), ERIC, PsychInfo, Education Full Text (EFT), and Educational Research Complete (ERC). Search terms were "rubric" AND ("validity" OR "achievement" OR "self-regulation" OR "student learning"). Additional articles were added from the references in these articles, making a total of 63 studies in this review. Criteria for selection were as follows:

(1) The article had to report on an original study and include some empirical data. Presentations of how to create and use rubrics, written for educator professional development, were excluded. Types of studies varied widely and included literature reviews, qualitative and quantitative studies of various designs, mixed method studies, instrument development studies, and essay case studies.
(2) The study had to be about descriptive rubrics that fit the definition used in this article, requiring both criteria (or "traits" or "dimensions") and performance level descriptions for each criterion. Studies whose "rubrics" were really rating scales, with criteria but no performance level descriptions, or some other type of point scheme for grading, were not retained.

Both authors read all studies and discussed decisions for both article selection and study description until consensus was reached

Findings

Research question 1, about the quality of assessment information from rubrics, is addressed in two sections, about reliability and validity, respectively. Research question 2, about the effects of rubric use on student learning and motivation to learn, is addressed in a third section.

Reliability

Thirty-eight of the studies reported one or more measures of reliability. Different measures of reliability measure consistency across different factors (e.g. raters, occasions, criteria), with different definitions of consistency (absolute, relative), using different calculation methods. Table 1 organizes reliability evidence by sorting studies according to the measures of reliability reported.

What measures of reliability of rubrics have been studied?

Table 1 shows that a wide variety of reliability measures have been used, and some studies reported more than one measure. While each of these measures reports some kind of consistency, they each give somewhat different information. A question thus arises: what are appropriate measures of reliability for rubrics? For judging the quality of student performance, consistency among raters is paramount. Measures of absolute agreement among raters are appropriate for the common educational model of teaching to standards, objectives, or course goals. Percent of rater agreement, Cohen's kappa (adjusting percent of agreement by the percentage expected by chance), intraclass correlation (ICC) for a single judge where judges are treated as a random factor, and generalizability coefficients for absolute decisions (although no study used this one) report amounts of absolute agreement. Measures of relative agreement among raters are appropriate for many research and evaluation purposes, for example studies of the relative effectiveness of an educational intervention. Pearson correlations, ICC for a single judge where judges are treated as a fixed factor, and generalizability coefficients for relative decisions report amounts of relative agreement.

Do rubrics reach commonly accepted reliability thresholds?

Appropriate levels of reliability differ according to purpose, too. For decisions about individuals (e.g. feedback, course grades), higher reliability is required than for decisions about groups (e.g. program evaluation). Nevertheless, it is instructive to compare the values in Table 1 with commonly accepted reliability thresholds. Graham, Milanowski, and Miller (2012) reported ranges of values generally considered acceptable: 75% to 90% agreement, Cohen's kappa of 0.61 to 0.81, and ICCs of 0.80 to 0.90. A common Cronbach's alpha threshold is 0.80 (Norcini 1999). However, other values are found in the literature, depending on the purpose of score use. For example, Iacobucci and Duhachek (2003) suggested alpha values of 0.70 can be acceptable for the early stages of research. Fleiss (1981, 1986) considered ICC values as low as 0.40 acceptable for some purposes, and characterized ICCs between 0.40 and 0.75 as in the fair to good range. On balance, Table 1 suggests that rubrics yield reliable results, or at least can, when criteria and performance level descriptions are clear and focused and when raters are trained.

Table 1. Reliability evidence for rubrics.

Study	Level	Rubric topic	Sample	Reliability evidence
Percent of Rater Agreement				
Garcia-Ros (2011)	Undergraduate	Oral presentation	64 educational psychology students	exact agreement = 66% adjacent agreement = 98%
Mansilla et al. (2009)	Undergraduate	Interdisciplinary writing	40 interdisciplinary essays plus some disciplinary essays	Four raters, exact agreement = 84%
Pagano et al. (2008)	Undergraduate	Writing (College composition)	Six institutions Year 1, five institutions Year 2	Adjacent agreement = 74%
Reddy (2011)	Graduate	Business cases, business projects	35 instructors, 95 business students, two institutions	Exact agreement 0.61–0.99
Stellmack et al. (2009)	Undergraduate	Writing APA-style introductions	40 papers, three researcher/graders	Interrater agreement exact = 0.37, adjacent = 0.90 intrarater agreement exact = 0.78, adjacent = 0.98
Wallace, Prather, and Duncan (2011)	Undergraduate	Astronomy − Cosmology	65 responses from 21 students, nine items	Exact agreement, overall score = 83%
Cohen's kappa				
Avanzino (2010)	Undergraduate	Oral communication	230 speeches (112 individual, 118 group)	$\kappa = 0.92$
Chasteen et al. (2012)	Undergraduate	Physics, electromagnetism	103 students in three courses (final version), 432 students in 14 courses during test development	$\kappa = 0.41$
Garcia-Ros (2011)	Undergraduate	Oral presentation	64 educational psychology students	$\kappa = 0.36$ exact agreement $\kappa = 0.80$ adjacent agreement

(*Continued*)

Table 1. (*Continued*).

Study	Level	Rubric topic	Sample	Reliability evidence
Panadero, Alonso-Tapia, and Reche (2013)	Undergraduate	Multimedia	69 pre-service teachers	$\kappa = 0.89$ and 0.92 (two rubrics)
Stellmack et al. (2009)	Undergraduate	Writing APA-style introductions	40 papers, three researcher/graders	$\kappa = 0.33$
Wallace, Prather, and Duncan (2011)	Undergraduate	Astronomy − Cosmology	65 responses from 21 students, nine items	$\kappa = 0.76$, weighted $\kappa = 0.82$
Pearson Correlations				
Reznitskaya et al. (2009)	Elementary	Argumentative reasoning	Compositions of 127 elementary students	Median $r = 0.78$
Schamber and Mahoney (2006)	Undergraduate	Critical thinking	2002, 30 papers; 2003, 30 papers	Median $r = 0.90$
Garcia-Ros (2011)	Undergraduate	Oral presentation	64 educational psychology students	Median $r = 0.89$
Intra-class correlations (ICCs)				
Cho, Schunn, and Wilson (2006)	Undergraduate, graduate	Writing	708 students in 16 courses over three years from four universities	Untrained raters single rater ICCs 0.17–0.56 multiple rater ICCs 0.45–0.88
De Wever et al. (2011)	Undergraduate	Group work	659 students in two years, in groups of 8–9 (81 groups)	Untrained raters single rater ICCs 0.33–0.50 (individual criteria), 0.50–0.60 (total score)
Kocakülah (2010)	Undergraduate	Newton's Laws of Motion problem solving	153 physics students in four classes	Untrained raters single rater ICCs 0.14, 0.38 multiple rater ICCs 0.93, 0.98
Lewis, Stiller, and Hardy (2008)	Undergraduate	Acute care treatment planning	22 students, five clinical educators, one academic faculty	Expert raters single rater ICC = 0.32

(*Continued*)

Table 1. (*Continued*).

Study	Level	Rubric topic	Sample	Reliability evidence
Newman et al. (2009)	Graduate faculty	Peer assessment of teaching	14 resource faculty	Expert raters single rater ICC = 0.27 (total score)
Nicholson, Gillis, and Dunning (2009)	Undergraduate	Nurse clinical performance in operating suite	40 pre-op nurses rating three videos	Expert raters single rater ICCs 0.51–0.61 multiple rater ICC = 0.98
Reddy (2011)	Graduate	Business cases, business projects	35 instructors, 95 business students, two institutions	Expert raters single rater ICCs 0.90–0.95 multiple rater ICCs 0.71–0.99
Rochford and Borchert (2011)	Graduate	Business case analysis	Case analysis assignments in MBA program capstone course	Expert raters multiple rater ICC = 0.96
Schreiber, Paul, and Shibley (2012)	Undergraduate	Public speaking competence rubric	Study 1, five coders, 45 speeches; Study 2, three undergraduate + one faculty coder, 50 speeches	Expert raters multiple rater ICCs 0.91, 0.93
Wald et al. (2012)	Graduate	Reflective writing	10−60 narratives over five trials	Expert raters single-rater ICCs 0.51–0.75
Generalizability coefficient				
Timmerman et al. (2011)	Undergraduate	Science writing	142 laboratory reports, nine trained and eight "natural" graduate student raters	Generalizability for relative decisions = 0.85
Cronbach's alpha				
Chasteen et al. (2012)	Undergraduate	Physics, electromagnetism	103 students in three courses (final version), 432 students in 14 courses during test development	Consistency between criteria $\alpha = 0.82$

(*Continued*)

Table 1. (*Continued*).

Study	Level	Rubric topic	Sample	Reliability evidence
Kocakülah (2010)	Undergraduate	Newton's Laws of Motion problem solving	153 physics students in four classes	Instructor's consistency between two forms, median $\alpha = 0.76$
Ciorba and Smith (2009)	Undergraduate	Music – instrumental and vocal performance	28 panels of judges, 359 music students' performances	Inter-judge consistency, median $\alpha = 0.89$
Wald et al. (2012)	Graduate	Reflective writing	10−60 narratives over five trials	Inter-judge consistency, median $\alpha = 0.77$
Studies using other methods				
Meier, Rich, and Cady (2006)	Middle school	Mathematics	Five teachers, 142 pieces of student work, 20 tasks	Describe the nature of disagreements between teachers and researchers' scores
Cho, Schunn, and Wilson (2006)	Undergraduate, graduate	Writing	708 students in 16 courses over three years from four universities	Compared reliability from student and instructor perspectives

Other ways of investigating reliability

Two of the studies did additional investigations of the question of how consistent scoring is when rubrics are used. One study (Meier, Rich, and Cady 2006) is described in this section. The other (Cho, Schunn, and Wilson 2006) is described in the section on Validity later. These studies are described in some detail because they expand arguments for reliability and validity beyond simply reaching a statistical threshold level.

Meier, Rich, and Cady (2006) studied a sample of five eighth-grade teachers who selected mathematics performance tasks to use in their classroom. Reliability was investigated by charting the amount of difference between the teachers' and researchers' scores. One teacher agreed with the researchers all the time. Two teachers disagreed with the researchers about half the time, most often when the problems students solved were familiar in mathematical content and required an equal mix of computation and explanation, and least on tasks that focused primarily on explanation. Two teachers disagreed with the researchers all the time and, while they only chose tasks that focused on familiar content, they too were more consistent using rubrics with problem-solving tasks that focused on explanation rather than computation.

Validity

Table 2 describes the approaches taken by studies in this review that reported validity evidence beyond the evidence of reliability reported in the previous section. All validity evidence is ultimately construct validity evidence; however, it is sometimes useful to distinguish various types of evidence that can be brought to bear on a validity argument. Table 2 shows that a wide variety of methods has been brought to bear on the question of the validity of scores from rubrics, and that in general the evidence has supported the validity of rubrics scores, with a few exceptions.

Content-related validity evidence

Content-related validity evidence has included documenting the source of rubric content (e.g. literature review, student work, course learning outcomes or standards) and expert review of the rubrics, although the experts were sometimes the same faculty members who developed the rubric. Such reviews were described for oral communication (Avanzino 2010), concepts of electrostatics (Chasteen et al. 2012), physical education (Dyson et al. 2011), media-enhanced science presentations (Mott et al. 2011), business education (Reddy 2011), science writing (Timmerman et al. 2011), and reflective writing (Wald et al. 2012).

Criterion-related validity evidence

Criterion-related validity evidence has included correlations of rubric scores with external judgments of the same work. Several studies reported validity as agreement of self or peer use of rubrics with the teacher or instructor's scores (Cho, Schunn, and Wilson 2006; Sadler and Good 2006). Schreiber, Paul, and Shibley (2012) presented evidence for the validity of scores on a public speaking rubric in the form of correlations between the rubric scores and the grades students had received on the

Table 2. Validity evidence for rubrics.

Study	Level	Rubric topic	Sample	Validity evidence
Avanzino (2010)	Undergraduate	Oral communication	230 speeches (112 individual, 118 group)	Based on student learning outcomes; subject expert review
Bauer and Cole (2012)	Undergraduate	Chemistry guided-inquiry activities	60 science faculty, four manipulated versions of the task	Rubric was sensitive enough to distinguish four versions of the activity
Chasteen et al. (2012)	Undergraduate	Physics, electromagnetism	103 students in three courses (final version), 432 students in 14 courses during test development	Expert feedback; student interviews. Student results differed by course (could differentiate types of instruction), criterion-related evidence (to physics grades)
Cho, Schunn, and Wilson (2006)	Undergraduate, graduate	Writing	708 students in 16 courses over three years from four universities	Correlations of student ratings with instructor and expert ratings
Ciorba and Smith (2009)	Undergraduate	Music – instrumental and vocal performance	28 panels of judges, 359 music students' performances	Scores rose by year (Fr-Soph-Jr-Sr); scale inter-correlations (internal validity evidence)
Doktor and Heller (2009)	Undergraduate	Written solutions to physics problems	Eight interviews (of 238 enrolled)	Think-aloud interviews
Dyson et al. (2011)	Elementary	Physical education – rubrics for tasks keyed to national standards	773 students, K, 2, 5; 20 project administrators	Content expert review
Garcia-Ros (2011)	Undergraduate	Oral presentation	64 educational psychology students	Students' perceptions
Hancock and Brundage (2010)	Graduate	Graduate student development profile for speech-language pathology students	Pilot 26 first year students, then applied whole-program	Demonstrated student growth over time; faculty perceptions
Hay and Macdonald (2008)	Secondary	Physical education (PE)	Two PE teachers, one from high socio-economic status (SES) and one low SES high school	[Teacher interviews, teachers did not always use rubrics but rather an intuitive assessment model – evidence did not support validity]

(*Continued*)

Table 2. (*Continued*).

Study	Level	Rubric topic	Sample	Validity evidence
Kocakülah (2010)	Undergraduate	Physics – Newton's Laws of Motion problems	153 physics students in four classes	Students' mean peer scores were same as instructor scores
Mansilla et al. (2009)	Undergraduate	Interdisciplinary writing	40 interdisciplinary essays plus some disciplinary essays	Scores successively higher across years (Fr/Soph–Sr); scores distinguished between disciplinary writing and interdisciplinary writing; rubric clear enough to diagnose student learning at a level of granularity sufficient to enable further instructional support
Moni, Beswick, and Moni (2005)	Undergraduate	Concept maps − Physiology	62 students, two faculties (plus one faculty advisor)	Student perceptions; faculty perceptions
Moni and Moni (2008)	Graduate	Concept maps − Physiology	61 dentistry students in groups of four, two of which were videotaped	Videotapes of student groups' use of rubrics; student perceptions
Mott et al. (2011)	Elementary, secondary	Media-enhanced science presentation rubric	Five experts, teacher education science method students	Expert review
Pagano et al. (2008)	Undergraduate	Writing (college composition)	Six institutions Year 1, five institutions Year 2	Scores increased from early to late in the semester
Reddy (2011)	Graduate	Business cases, business projects	35 instructors, 95 business students, two institutions	Expert review; student perceptions
Rezaei and Lovorn (2010)	Graduate	Writing	467 graduate students	Quasi-experiment investigating influence of construct-irrelevant factors
Reznitskaya et al. (2009)	Elementary	Argumentative reasoning	Compositions of 127 elementary students	Compare analytic and holistic versions (holistic only captured some criteria cf. analytic); factor analysis (internal structure evidence)

(*Continued*)

Table 2. (*Continued*).

Study	Level	Rubric topic	Sample	Validity evidence
Sadler and Good (2006)	Middle school	Task-specific rubrics for constructed response science test items	95 students in four general science classrooms, one teacher	Criterion-related evidence (students'/ teachers' ratings correlations)
Schreiber, Paul, and Shibley (2012)	Undergraduate	Public speaking competence rubric	Study 1, five coders, 45 speeches; Study 2, three undergraduate + one faculty coder, 50 speeches	Factor analysis (internal structure evidence); criterion-related evidence (correlation of rubric scores for speeches with grades assigned to the speeches using different scoring schemes during the semester)
Spence (2010)	Elementary	Writing	Two third grade teachers, one ELL student	Observations; interviews, think-alouds
Stellmack et al. (2009)	Undergraduate	Writing APA-style introductions	40 papers, three researcher/ graders	Criterion-related evidence (Spearman correlation with independent judge)
Timmerman et al. (2011)	Undergraduate	Science writing	142 laboratory reports, nine trained and eight "natural" graduate student raters	Grader (graduate student) perceptions; faculty (expert) review
Wald et al. (2012)	Graduate	Reflective writing	10−60 narratives over five trials	Rubric content based on literature
Wallace, Prather, and Duncan (2011)	Undergraduate	Astronomy − Cosmology	65 responses from 21 students, nine items	Rubric content based on student responses to tasks

same speeches in their classroom (graded by different grading schemes than the rubric). Kocakülah (2010) found that instructors', peers', and an independent coder's scores on a rubric for evaluating students' problem-solving with Newton's Laws of Motion did not differ from one another. Stellmack et al. (2009) studied the correlation between an independent judge's rankings of APA-style introductions in a research method class and those obtained with the rubric.

Internal validity evidence

The purpose of internal validity evidence is to demonstrate the relationships among the criteria within a rubric. Internal validity evidence has included factor analyses (Reznitskaya et al. 2009; Schreiber, Paul, and Shibley 2012) and scale inter-correlations (Ciorba and Smith 2009). Reznitskaya et al. (2009) factor-analyzed analytic rubrics for scoring argumentative reasoning. They re-analyzed data comparing two different instructional methods for developing students' argumentation skills, comparing results of using the factor scores and using a holistic rubric. Using the two factors, the treatment produced statistically significant differences in argumentation; using the holistic scores, the treatment effect was not significant. The researchers interpreted this to mean the analytic rubric provided a fuller measure of the construct.

Other construct-related evidence for validity

Other construct-related validity evidence has included demonstrating that scores behaved as predicted. Several studies investigated predicted improvement over time and with instruction (Ciorba and Smith 2009; Hancock and Brundage 2010; Mansilla et al. 2009; Pagano et al. 2008; Wallace, Prather, and Duncan 2011).

Bauer and Cole (2012) demonstrated that scores on a rubric for evaluating desired characteristics in chemistry activities was sensitive enough to capture differences among four versions of an experimentally-manipulated nuclear atom activity. Mansilla et al. (2009) showed that an interdisciplinary writing rubric was sensitive enough to distinguish between interdisciplinary and discipline-based writing. Docktor and Heller (2009) studied student interviews to establish that students did use the various characteristics of problem-solving enumerated as criteria in a physics problem-solving rubric as they solved the problems.

One study did not support the validity of rubric scores. Rezaei and Lovorn (2010) found that using a rubric to assess writing in college social studies increased the range of assigned scores to a given essay, increasing the standard error of measurement, and that student graders were strongly influenced by the mechanics of the writing.

Consequential evidence for validity

Consequential evidence for validity has included perceptions of students, faculty, and/or teaching assistants (Garcia-Ros 2011; Moni, Beswick, and Moni 2005; Moni and Moni 2008) and analyses of videotapes of students' use of rubrics (Moni and Moni 2008). Most often rubrics were deemed useful and helpful. However, Hay and Macdonald (2008) found, using interviews, that many school teachers did not use physical education rubrics, but rather evaluated their students "intuitively." This

result, of course, is evidence for lack of validity in that scores represented global judgments and included idiosyncratic criteria rather than reflecting the criteria and performance descriptions in the rubric. Spence (2010) found that third grade teachers using the six Traits writing rubric with English Language Learners (ELLs) sometimes used the rubric too rigidly, sometimes not crediting ELL students' writing as demonstrating these traits when, given their context, it probably did.

Different perspectives on the reliability and validity of rubrics

Cho, Schunn, and Wilson (2006) studied the validity and reliability of scaffolded peer assessment of writing in a university setting from two perspectives, distinguishing between the instructor and student points of view. They reasoned that instructors and students would have different concerns about peer evaluation. Instructors would want evidence of reliability and validity of peer assessment in the conventional sense, in order to be persuaded that the evaluations were accurate and useful. Also, they pointed out, instructors have access to the full range of student work and have a "macro view" (892) of the validity of ratings of student work. Therefore, to investigate validity from the instructor perspective, they calculated Pearson correlations of mean peer ratings with instructor ratings, which treat instructor ratings as the true score. To investigate reliability from the instructor perspective, they reported ICCs for one and for multiple peer raters.

Cho, Schunn, and Wilson (2006) reasoned that students might have a different perspective because of their "micro view" (892): students are concerned mostly with their own papers and do not have access to all students' work. Students might think of validity in terms of receiving peer reviews that were close to instructor ratings, and of reliability in terms of having a small amount of variability (or in other words, a large amount of agreement) in peer ratings of their own work. To investigate validity from the student perspective, Cho, Schunn, and Wilson (2006) measured distance between peer and instructor rating. To investigate reliability from the student perspective, they reported mean standard deviations in peer ratings per paper. These were often greater than the class standard deviation. The researchers concluded that peer ratings (of at least four peers) were both highly reliable and as valid as the instructor ratings, while paradoxically seeming unreliable and invalid to the students.

Utility

Utility is not the same as validity, but it is a characteristic that has come to be considered important for assessment and evaluation (Joint Committee on Standards for Educational Evaluation 2003). People will not have confidence in an assessment method not perceived as useful, and will likely not use the scores resulting from it. Like validity, utility also is referenced to purpose ("useful for what?").

Seventeen of the studies (Andrade et al. 2009; Bissell and Lemons 2006; Dinur and Sherman 2009; Fraser et al. 2005; Gerretson and Golson 2005; Green and Bowser 2006; Harnden 2005; Kerby and Romine 2010; Knight 2006; Loeffler 2005; McCormick, Dooley, Lindner, and Cummins 2007; Pagano et al. 2008; Peach, Mukherjee, and Hornyak 2007; Petkov and Petkova 2006; Rochford and Borchert 2011; Schlitz et al. 2009; Siegel et al. 2011) included in this review were categorized as "essay case studies." These were show-and-tell pieces to testify to readers what

Table 3. Studies of the effects of rubric use on student learning and motivation to learn.

Study	Level	Rubric topic	Sample	Design[1]	Findings
Andrade and Du (2005)	Undergraduate	Educational Psychology	14 teacher education students who had used rubrics in Ed Psych	Focus groups	Students used rubrics to determine teacher's expectations, plan production, check their work in progress, and guide and reflect on feedback. Some students only checked the A and B levels of the rubric, and some saw rubrics as a way to "give teachers what they want."
Andrade, Du, and Mycek (2010)	Middle	Writing, 6+1 trait	162 students in 11 classes taught by six teachers	Quasi-experimental	Main effects of treatment (using rubrics), gender (girls), grade level, writing time, previous achievement in English, and for all seven criteria in the rubric. Using rubrics led to more effective writing.
Andrade, Du, and Wang (2008)	Elementary	Writing, 6+1 trait	116 students in seven classes	Quasi-experimental	Main effects of treatment (using rubrics), after controlling for previous achievement in English, and for four criteria in the rubric (Ideas, Organization, Paragraphs, Voice, Word choice). Using rubrics led to more effective writing.
Andrade et al. (2009)	Elementary, middle	Writing, 6+1 trait	268 students in 18 classes	Quasi-experimental	Self-efficacy rose more for the treatment (rubrics) group but not significantly so. Self-efficacy for the comparison group rose too as they worked through the writing process. Self-efficacy for girls was related to short-term rubric use. Only partial support for the claim that using rubrics for self-assessment raises self-efficacy.
Ash, Clayton, and Atkinson (2005)	Undergraduate	Service learning objectives, Critical thinking	14 students in two classes	Pre-experimental	Improvement across drafts was noted, with the Academic criterion being the most difficult for students. Improvement in first drafts across the semester was also noted, but smaller, and again the Academic criterion was the hardest.

(*Continued*)

Table 3. (*Continued*).

Study	Level	Rubric topic	Sample	Design[1]	Findings
Coe et al. (2011)	Elementary	Writing, 6+1 trait	74 schools, TRT = 102 teachers, 2230 students; CNTRL = 94 teachers, 1931 students	Experimental	After controlling for baseline achievement and school and teacher characteristics, treatment group (whose teachers had professional development in using the rubrics) outscored the control group by 0.109 standard deviation. For three traits (Organization, Voice, Word choice) the difference was significant, for the other three the difference was in the same direction but not significant.
Howell (2011)	Undergraduate	Juvenile delinquency course assignment rubric	80 students in two sections of the instructor's own course	Quasi-experimental	Controlling for college year, criminal justice major (vs. not), pre-test score and gender, being in the treatment group (having rubrics provided with the assignment) predicted achievement ($\beta = 0.488$). The only other large predictor was college year. Student achievement was higher when rubrics were used.
Jonsson (2010)	Undergraduate	Teacher education	2004, 170 students; 2005, 154 students; 2006, 138 students	Pre-experimental	Changes in transparency (adding self-assessment, rubrics, and exemplars to an exam), resulted in significantly higher performance
Kocakülah (2010)	Undergraduate	Newton's Laws of Motion problem solving	153 physics students in four classes	Quasi-experimental	Students who took part in the designing and using of a rubric, performed better in solving problems than those who had the same instruction but no rubric.
Lee and Lee (2009)	Grades 5 and 6 special education	Classroom engagement behaviors	Three Grade 5 or Grade 6 special education students, MMR	Single subject	Target behaviors increased from baseline, after intervention (rubrics) for all three students: Listen to lecture, take notes, work with peers, work independently.

(*Continued*)

Table 3. (*Continued*).

Study	Level	Rubric topic	Sample	Design[1]	Findings
Panadero, Alonso-Tapia, and Reche (2013)	Undergraduate	Multimedia	69 pre-service teachers	Quasi-experimental	Using the scripts resulted in higher levels of self-regulation. Rubrics decreased performance/avoidance self-regulation (negative self-regulatory actions detrimental to learning). No significant effects for students' performance or self-efficacy. Students preferred the use of rubrics to the use of scripts.
Panadero, Tapia, and Huertes (2012)	Secondary	Social studies, analysis of landscapes	120 secondary school students	Experimental	Rubrics for self-assessment were compared with scripts (focused questions) and with using no tool. Students who used scripts had the highest self-regulation scores, followed by rubrics, and then control. Students who used rubrics and scripts both outperformed the control group in achievement.
Reynolds-Keefer (2010)	Undergraduate	Writing	45 Ed psych students	Open-ended questionnaire	Pre-service teachers who used rubrics as students reported being more likely to use rubrics in their own teaching.
Sadler and Good (2006)	Middle school	Science test items	95 students in four general science classrooms, one teacher	Quasi-experimental	Students who self-graded their own science tests with a rubric scored significantly higher than students who graded peers' tests with rubrics and than a control group, on an unannounced, second administration of the test a week later.
Vandenberg et al. (2010)	Undergraduate	Financial analysis project	49 students in three sections of the course	Pre-experimental	Students who used rubrics scored significantly higher on two of three sections of the project. Students with rubrics felt the requirements of the assignment were more clearly communicated that those without.
Yopp and Rehberger (2009)	Undergraduate	Pre-algebra	32 students in four sections, two instructors	Quasi-experimental	Treatment (which included focused feedback as well as rubric use) group scored significantly higher than control on final exam and mathematics self-efficacy.

[1]Quantitative designs are categorized according to Campbell and Stanley's (1963) categories: pre-experimental, quasi-experimental, and experimental.

had been done with rubrics and in what ways the rubrics were useful. Many presented "lessons learned" and were intended as models and, sometimes, inspirations for other faculty who might be interested in developing rubrics for their courses or programs. Most (13) were in higher education; all testified to the importance and usefulness of rubrics. It is worth noting that most of these studies were written from the faculty perspective and had very little to say about the student perspective, a point which will be taken up in the discussion.

Effects of rubric use on student learning and motivation to learn

Table 3 shows the evidence for effects on student learning and motivation as reported in 16 studies from this review. Teachers, not students, were the subjects in two other studies in this category, in science (Harnden 2005) and the arts (Mason, Steedly, and Thormann 2008). The 16 studies in Table 3 included two experimental, eight quasi-experimental, three pre-experimental, and one single-subject quantitative research designs, and two qualitative designs which used focus groups and an open-ended questionnaire, respectively. The quantitative designs tested hypotheses about increased student performance, self-regulation or self-efficacy, or both, with rubric use. The qualitative designs reported participants' perceived changes after rubric use.

Rubrics and performance

Thirteen of the studies addressed questions about whether achievement or performance rose for groups that used rubrics. Rubric use was associated with increased student achievement in writing (Andrade, Du, and Mycek 2010; Andrade, Du, and Wang 2008; Coe et al. 2011), general science (Sadler and Good 2006), social studies (Panadero, Tapia, and Huertes 2012), physics (Kocakülah 2010), mathematics (Yopp and Rehberger 2009), service learning (Ash, Clayton, and Atkinson 2005), business education (Vandenberg et al. 2010), criminal justice (Howell 2011), teacher education (Jonsson 2010), and special education (Lee and Lee 2009; Loeffler 2005). One study (Panadero, Alonso-Tapia, and Reche 2013) showed no significant effects of rubric use on performance.

Two of these studies will be described in more detail because of the rigor of their designs. Coe et al. (2011) conducted the only study in this review whose sample was not institutionally bound and used an experimental design, including random selection of participants. Panadero, Tapia, and Huertas (2012) also used an experimental design and investigated effects of rubric use on self-regulation as well as on achievement.

Coe et al. (2011) investigated the impact of the 6 + 1 Trait Writing Model on student writing achievement in Grade 5. The 6 + 1 Trait Writing Model is centered in a set of rubrics that are widely used in writing assessment and instruction in elementary and secondary settings in the United States and around the world (available from http://educationnorthwest.org/traits). The "traits" are the rubric's six criteria for writing quality (ideas, organization, voice, word choice, sentence fluency, conventions) plus an optional criterion called presentation for use when a polished, visually appealing final product is required.

The cluster-randomized experimental study collected data from 74 schools in Oregon, over a period of two years. Students wrote essays at the beginning and the

end of the school year in which they participated. Mean difference between pre- and post-essay scores were compared in a statistical model that took into account the nested nature of the data and controlled for baseline writing performance, school poverty level, and school averages for weekly teacher-reported hours students spend in class practicing writing, years of teacher experience, and years of teacher experience teaching writing. Overall, the estimated (after controls) average score of students in the treatment group was higher than the estimated average score of students in the control group.

Panadero, Tapia, and Huertas (2012) conducted their study in geography classes in two secondary schools in Spain, using the lens of self-regulation. The three between-group independent variables were: (1) type of instruction (oriented to process or to performance), (2) type of self-assessment tool (control vs. rubric vs. script), and (3) feedback (oriented to process or to performance). There was also one within-group variable, task number (first through third). The tasks were analyses of landscapes, a usual task in the secondary geography classes. The self-assessment factor compared students who used a rubric to assess their work with students who used a script (a series of self-reflection questions that walked students through the process of doing the task) and with a control group that used neither tool. The rationale was that rubrics would focus students more on the product, the completed landscape analysis, and scripts would focus students more on the process of doing the task. Students who used rubrics or scripts out-performed the control group, suggesting either tool fostered learning. Students who used scripts had the highest self-regulation scores, followed by rubrics, and then control. There was an occasion effect, with the highest self-regulation noted at the first task, decreasing over the second and third tasks.

Rubrics and self-regulation of learning

Four of the studies addressed questions about whether using rubrics was associated with gains in self-regulation of learning or in self-efficacy, which is one of the constellation of variables usually included in theories of self-regulation of learning (Pintrich and Zusho 2002; Zimmerman 2011). Three of these (Panadero, Tapia, and Huertas 2012; Panadero, Alonso-Tapia, and Reche 2013; Yopp and Rehberger 2009) investigated effects on achievement as well as on these motivational variables. Using rubrics is associated with increased student self-efficacy in elementary and middle school writing (Andrade et al. 2009) and in undergraduate mathematics (Yopp and Rehberger 2009). Using rubrics is associated with increased student self-regulation in secondary social studies (Panadero, Tapia, and Huertas 2012) and in undergraduate teacher education (Panadero, Alonso-Tapia, and Reche 2013).

However, the evidence is mixed within these studies. Each study demonstrated that rubrics have positive effects on motivation, but not in all cases and with all measures. The study by Panadero, Tapia, and Huertas (2012) was described earlier. Andrade et al. (2009) investigated the relationship between long- and short-term rubric use, gender, and self-efficacy for writing. Girls' self-efficacy was higher than boys' self-efficacy before they began writing. Average self-efficacy ratings increased as students wrote, regardless of condition, but the increase in the self-efficacy of girls in the treatment group was larger than that for girls in the comparison group, and long-term rubric use was associated only with the self-efficacy of girls.

Panadero, Alonso-Tapia, and Reche (2013) investigated the effects of rubrics and scripts among pre-service teachers learning to design multimedia materials. Students using the scripts had higher levels of self-regulation for learning. Students using rubrics decreased in performance/avoidance self-regulation, that is, in negative self-regulatory actions that would harm learning. There were no significant effects on performance or self-efficacy.

Rubrics and student attitudes and perceptions

Three studies, all at the undergraduate level, investigated students' perceptions about the effects of rubrics on their work. Andrade and Du (2005) studied the attitudes and experiences of teacher education students who had used rubrics in their educational psychology class. Many of the students reported they used rubrics to determine the teacher's expectations, plan their work as they completed it, self-assess their work in progress, and reflect on feedback they received. Some students reported they only looked at the descriptions of performance for the A and B levels on the rubric, so that they could "give teachers what they want" (4). Reynolds-Keefer (2010) also studied pre-service teachers in educational psychology classes. Similarly to Andrade and Du (2005), the students reported using the rubrics to understand teacher expectations, but most of their description of expectations concerned the number of points for various attributes of the work that would contribute to the final grade. They did not consider the rubric as a tool for reflection. Many did, however, report that they would be more likely to use rubrics in their own future teaching.

Vandenberg et al. (2010) tested differences in student attitudes and perceptions, measured quantitatively. Forty-nine students in three sections of a financial accounting course completed a group financial analysis project. Two of the sections used a rubric for this project, and one did not. There were no significant differences between students who did and did not use rubrics on questions about the clarity of the learning objective, the clarity of the writing requirements for the project, the clarity of the information gathering requirements for the project, or the clarity of requirements for presenting financial data. Students without the rubric reported feeling clearer about requirements relating to synthesizing financial and non-financial data for a conclusion. Students without the rubric reported exerting more effort. Despite these mostly non-significant findings in student perceptions, students who used the rubric scored significantly higher on parts one and three of the three-part project. Thus it seems there is not a straight line between students' perceptions and their performance.

Discussion

This review showcases a literature that is beyond its infancy but not yet mature. Overall, the rigor and scope of the studies has increased since the first literature review on rubrics (Jonsson and Svingby 2007). Regarding reliability, the sophistication of the measures used has increased, most notably with an increase in the number and type of ICCs reported. Regarding validity, the amount and range of evidence has increased. Regarding the effects of rubric use on learning and motivation, some studies featured experimental designs. The following sections summarize findings and suggest implications for future research.

Reliability

This review has demonstrated that raters using rubrics can achieve acceptable levels of consistent and reliable judgment, even though they do not always do so. When rubrics do not reach acceptable levels of reliability, several explanations appear plausible. The clarity of the rubrics themselves and the level of expertise, training, and investment of the raters are the two most obvious factors to consider. Future reliability research might evaluate and take into account the quality of the rubrics as an example of the genre, as judged by an expert in rubric design and writing. Future reliability research might also take into account levels of faculty training and investment in rubric design. Using rubrics well, like doing anything well, requires training careful attention. Finally, future reliability research should: (1) justify the choice of reliability measures in light of the purpose for which the rubric scores will be used; (2) report clearly whether the researchers are interested in absolute or relative agreement among raters (sometimes called "agreement" or "consistency") – as some of the studies did – and (3) justify the choice based on the intended use for the rubric's information. For most purposes related to student learning, the appropriate choice will be absolute agreement, which is a more stringent test to meet than relative agreement.

Validity

While the validity evidence shown in the total body of studies summarized in this review is impressive in its use of a variety of methods, on balance there is more work to be done. Most of the studies used only one or two methods of gathering validity evidence. Bias was an issue in many of them, for example, when the authors of the rubric or their colleagues or students furnished expert review, survey, or interview evidence.

Almost all of the evidence reviewed in the sections on reliability and validity of information from rubrics has been from post-secondary education settings. Many of the studies had as their major purpose establishing the reliability or validity of a rubric for faculty or program use, without too much regard for students using the rubric. Program accreditation requirements, plus the general tendency of higher education faculty to study and publish their work, may have contributed to this result. However, there seems to be enough evidence to conclude that rubrics can produce valid and useful scores for grading or program evaluation in post-secondary education. In fact, many scoring schemes could be used if all that is required is accurate and valid accountability information.

However, one of the major arguments for using rubrics has been that they are a useful tool for students to use in learning (Panadero and Jonsson 2013). Future research should investigate this formative use of rubrics through the lens of validity theory. The aim should be to provide a body of evidence for the validity of rubric use in formative assessment that is at least on a par with the current evidence for the validity of rubric use for summative assessment. Studies of the effects of rubrics on student learning and motivation provide consequential evidence for the validity of formative use of rubrics. Other types of validity evidence for the formative use of rubrics are not as plentiful.

Effects of rubric use on student learning and motivation

Regarding the effects of rubric use on learning and performance, this review found 13 studies, many of which used relatively rigorous designs, including two

experiments and eight quasi-experimental studies. The body of evidence that is accumulating is promising but not sufficient for establishing that using rubrics cause increased performance. The designs of these studies allow for many competing hypotheses. There are three main reasons for this, which may be addressed in future research.

First, only two of the studies in this review were experimental studies. Second, all the studies but one (Coe et al. 2011) were conducted with convenience samples. Generalization is limited to the schools, classes, or in some cases, individual teacher within which the sample was nested. Third, in many of the studies rubric use was confounded with other aspects of treatment (feedback, self-assessment, and so on). This confounding of rubrics and other instructional and/or formative assessment methods may reflect an important reality. It is probably not wise to make claims for rubrics per se, but for rubrics as a particularly user-friendly form of making the qualities of good work explicit (for formative assessment) and making final expectations explicit (for grading). Other authors have made this point (Jonsson 2010; Panadero, Tapia, and Huertas 2012; Torrance 2007). As a vehicle for communicating expectations and a tool for self-assessment, rubrics are not a "method" unto themselves. Rubrics necessarily need to be part of instructional and formative assessment methods. Future research should also then address the use of rubrics with various instructional and formative assessment strategies.

It may turn out that it is not rubrics per se (that is, rubrics as an assessment tool in a particular form), but the provision of focused learning goals, criteria, and performance descriptions in whatever form that supports learning and motivational outcomes for students. Even so, the fact that rubrics are an efficient, clear, and easily understood way to focus learning goals, criteria, and performance descriptions would recommend their use as one of, or even the primary, form in which to do this.

Regarding the effects of rubric use on motivation, the fact that four studies had generally positive but mixed effects suggests that there is more work to be done to describe the nature of the effects of rubrics on self-regulation of learning. It appears that other self-assessment tools (e.g. scripts) that focus students on the task also affect self-regulation of learning. There may be a gender effect, but whether it is specific to writing, where girls may have an advantage (Brookhart 2009) or more general is unknown. Self-regulation itself encompasses a large array of cognitive, motivational, behavioral, and contextual variables (Pintrich and Zusho 2002); study of the relationships among these variables and rubrics use is just beginning.

Implications for practice

The findings support the use of rubrics at all educational levels with the exception of early childhood, where no studies were found. In turn, this means teachers or post-secondary faculty should have professional development in using rubrics and in coaching students to use rubrics, and pre-service teachers should have training in these matters, as well.

Finally, readers should keep in mind that recommendations based on these findings depend on the rubrics being suited for both formative and summative use by containing description of quality work and not evaluation only. Although it is not recommended (Andrade 2000; Arter and Chappuis 2006; Brookhart 2013b), many rubrics do use "descriptions" of performance that read more like points-for-requirements (e.g. "has three sources") than quality of work (e.g. "consults and interprets

appropriate sources"). Rubrics that focus on the requirements for an assignment and not indications of learning, especially if used in classroom contexts where evaluation is more salient than learning (Stiggins and Conklin 1992), cannot be expected to support students' focus on learning over grading. Rubrics that include descriptions of quality on criteria that reflect learning goals, however, confer benefits that simple rating scales or point schemes cannot; they function as the goals toward which students can monitor their progress.

References

Andrade, Heidi Goodrich. 2000. "Using Rubrics to Promote Thinking and Learning." *Educational Leadership* 57 (5): 13–18.

Andrade, Heidi L., and Gregory J. Cizek, eds. 2010. *Handbook of Formative Assessment*. New York: Routledge.

Andrade, Heidi, and Ying Du. 2005. "Student Perspectives on Rubric-referenced Assessment." *Practical Assessment, Research & Evaluation* 10 (3). http://pareonline.net/pdf/v10n3.pdf.

Andrade, Heidi L., Ying Du, and Xiaolei Wang. 2008. "Putting Rubrics to the Test: The Effect of a Model, Criteria Generation, and Rubric-referenced Self-assessment on Elementary School Students' Writing." *Educational Measurement: Issues and Practice* 27 (2): 3–13. doi:10.1111/j.1745-3992.2008.00118.x.

Andrade, Heidi, Colleen Buff, Joe Terry, Marilyn Erano, and Shaun Paolino. 2009. "Assessment-driven Improvements in Middle School Students' Writing." *Middle School Journal* 40 (4): 4–12.

Andrade, Heidi L., Xiaolei Wang, Ying Du, and Robin L. Akawi. 2009. "Rubric-referenced Self-assessment and Self-Efficacy for Writing." *The Journal of Educational Research* 102 (4): 287–302.

Andrade, Heidi L., Du Ying, and Kristina Mycek. 2010. "Rubric-referenced Self-assessment and Middle School Students' Writing." *Assessment in Education: Principles, Policy & Practice* 17 (2): 199–214.

Arter, Judith A., and Jan Chappuis. 2006. *Creating & Recognizing Quality Rubrics*. Boston, MA: Educational Testing Service.

Ash, Sarah L, Patti H. Clayton, and Maxine P. Atkinson. 2005. "Integrating Reflection and Assessment to Capture and Improve Student Learning." *Michigan Journal of Community Service Learning* 11 (2): 49–60.

Avanzino, Susan. 2010. "Starting from Scratch and Getting Somewhere: Assessment of Oral Communication Proficiency in General Education across Lower and Upper Division Courses." *Communication Teacher* 24 (2): 91–110. doi:10.1080/17404621003680898.

Bauer, Christopher F., and Renée Cole. 2012. "Validation of an Assessment Rubric via Controlled Modification of a Classroom Activity." *Journal of Chemical Education* 89 (9): 1104–1108.

Bissell, Ahrash N., and Paula P. Lemons. 2006. "A New Method for Assessing Critical Thinking in the Classroom." *BioScience* 56 (1): 66–72.

Brookhart, Susan M. 2009. "Assessment, Gender and in/Equity." In *Educational Assessment in the 21st Century*, edited by Claire Wyatt-Smith and J. Joy Cumming, 119–136. Dordrecht: Springer.

Brookhart, Susan M. 2013a. "The Public Understanding of Assessment in Educational Reform in the United States." *Oxford Review of Education* 39: 52–71.

Brookhart, Susan M. 2013b. *How to Create and Use Rubrics for Formative Assessment and Grading*. Alexandria, VA: ASCD.

Campbell, Donald T., and Julian C. Stanley. 1963. *Experimental and Quasi-experimental Designs for Research*. Chicago, IL: Rand McNally.

Chasteen, Stephanie V., E. Rachel Pepper, Marcos D. Caballero, Steven J. Pollock, and Katherine K. Perkins. 2012. "Colorado Upper-division Electrostatics Diagnostic: A Conceptual Assessment for the Junior Level." *Physical Review Special Topics – Physics Education Research* 8 (2): 020108. doi:10.1103/PhysRevSTPER.8.020108.

Cho, Kwangsu, Christian D. Schunn, and Roy W. Wilson. 2006. "Validity and Reliability of Scaffolded Peer Assessment of Writing from Instructor and Student Perspectives." *Journal of Educational Psychology* 98 (4): 891–901. doi:10.1037/0022-0663.98.4.891.

Ciorba, C. R., and N. Y. Smith. 2009. "Measurement of Instrumental and Vocal Undergraduate Performance Juries Using a Multidimensional Assessment Rubric." *Journal of Research in Music Education* 57 (1): 5–15. doi:10.1177/0022429409333405.

Coe, Michael, Makoto Hanita, Vicki Nishioka, and Richard Smiley. 2011. *An Investigation of the Impact of the 6 + 1 Trait Writing Model on Grade 5 Student Writing Achievement* (NCEE 2012–4010). Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education.

De Wever, Bram, Hilde Van Keer, Tammy Schellens, and Martin Valcke. 2011. "Assessing Collaboration in a Wiki: The Reliability of University Students' Peer Assessment." *The Internet and Higher Education* 14 (4): 201–206. doi:10.1016/j.iheduc.2011.07.003.

Dinur, Adva, and Herbert Sherman. 2009. "Incorporating Outcomes Assessment and Rubrics into Case Instruction." *Journal of Behavioral and Applied Management* 10 (2): 291–311.

Docktor, Jennifer, and Kenneth Heller. 2009. "Assessment of Student Problem Solving Processes." *AIP Conference Proceedings* 1179: 133–136.

Dyson, Ben, Judith H. Placek, Kim C. Graber, Jennifer L. Fisette, Judy Rink, Weimo Zhu, and Marybell Avery, et al. 2011. "Development of PE Metrics Elementary Assessments for National Physical Education Standard 1." *Measurement in Physical Education and Exercise Science* 15 (2): 100–118. doi:10.1080/1091367X.2011.568364.

Fleiss, Joseph L., ed. 1981. *Statistical Methods for Rates and Proportions*. New York: John Wiley.

Fleiss, Joseph L. 1986. *The Design and Analysis of Clinical Experiments*. New York: John Wiley.

Fraser, Linda, Katrin Harich, Joni Norby, Kathy Brzovic, Teeanna Rizkallah, and Dana Loewy. 2005. "Diagnostic and Value-added Assessment of Business Writing." *Business Communication Quarterly* 68 (3): 290–305.

García-ros, Rafael. 2011. "Analysis and Validation of a Rubric to Assess Oral Presentation Skills in University Contexts." *Electronic Journal of Research in Educational Psychology* 9 (3): 1043–1062.

Gerretson, Helen, and Emily Golson. 2005. "Synopsis of the Use of Course-embedded Assessment in a Medium Sized Public University's General Education Program." *JGE: The Journal of General Education* 54 (2): 139–149.

Graham, Matthew, Anthony Milanowski, and Jackson Miller. 2012. *Measuring and Promoting Inter-Rater Agreement of Teacher and Principal Performance Ratings*. Center for Educator Compensation Reform. cecr.ed.gov/pdfs/Inter_Rater.pdf.

Green, Rosemary, and Mary Bowser. 2006. "Observations from the Field: Sharing a Literature Review Rubric." *Journal of Library Administration* 45 (1): 185–202. doi:10.1300/J111v45n01_10.

Hancock, Adrienne B., and Shelley B. Brundage. 2010. "Formative Feedback, Rubrics, and Assessment of Professional Competency through a Speech-language Pathology Graduate Program." *Journal of Allied Health* 39 (2): 110–119.

Harnden, Jodie. 2005. "Scientific Inquiry Scoring Guides." *Science Scope* 28 (4): 52–54.

Hay, Peter J., and Doune Macdonald. 2008. "(Mis)Appropriations of Criteria and Standards-referenced Assessment in a Performance-based Subject." *Assessment in Education: Principles, Policy & Practice* 15 (2): 153–168. doi:10.1080/09695940802164184.

Howell, Rebecca J. 2011. "Exploring the Impact of Grading Rubrics on Academic Performance: Findings from a Quasi-experimental, Pre-post Evaluation." *Journal on Excellence in College Teaching* 22 (2): 31–49.

Iacobucci, Dawn, and Adam Duhachek. 2003. "Advancing Alpha: Measuring Reliability with Confidence." *Journal of Consumer Pscyhology* 13 (4): 478–487.

Joint Committee on Standards for Educational Evaluation. 2003. *The Student Evaluation Standards: How to Improve Evaluations of Students*. Thousand Oaks, CA: Corwin Press.

Jonsson, Anders. 2010. "The Use of Transparency in the 'Interactive Examination' for Student Teachers." *Assessment in Education: Principles, Policy & Practice* 17 (3): 183–197.

Jonsson, Anders, and Gunilla Svingby. 2007. "The Use of Scoring Rubrics: Reliability, Validity and Educational Consequences." *Educational Research Review* 2 (2): 130–144. doi:10.1016/j.edurev.2007.05.002.

Kerby, Debra, and Jeff Romine. 2010. "Develop Oral Presentation Skills through Accounting Curriculum Design and Course-embedded Assessment." *Journal of Education for Business* 85 (3): 172–179.

Knight, Lorrie A. 2006. "Using Rubrics to Assess Information Literacy." *Reference Services Review* 34 (1): 43–55.

Kocakülah, Mustafa Sabri. 2010. "Development and Application of a Rubric for Evaluating Students' Performance on Newton's Laws of Motion." *Journal of Science Education and Technology* 19 (2): 146–164. doi:10.1007/s10956-009-9188-9.

Lane, Suzanne, and Sean T. Tierney. 2008. "Performance Assessment." In *21st Century Education: A Reference Handbook* (Vol. 1), edited by Thomas L. Good, 461–470. Los Angeles, CA: SAGE.

Lee, Eunjung, and Sohyun Lee. 2009. "Effects of Instructional Rubrics on Class Engagement Behaviors and the Achievement of Lesson Objectives Typical Peers." *Education and Training in Developmental Disabilities* 44 (3): 396–408.

Lewis, Lucy K., Kathy Stiller, and Frances Hardy. 2008. "A Clinical Assessment Tool Used for Physiotherapy Students-is It Reliable?" *Physiotherapy Theory and Practice* 24 (2): 121–134.

Loeffler, Kelly A. 2005. "No More Friday Spelling Tests? An Alternative Spelling Assessment for Students with Learning Disabilities." *Teaching Exceptional Children* 37 (4): 24–27.

Mansilla, Veronica Boix, Elizabeth Dawes Duraisingh, Christopher R. Wolfe, and Carolyn Haynes. 2009. "Targeted Assessment Rubric: An Empirically Grounded Rubric for Interdisciplinary Writing." *Journal of Higher Education* 80 (3): 334–353.

Mason, C. Y., K. M. Steedly, and M. S. Thormann. 2008. "Impact of Arts Integration on Voice, Choice, and Access." *Teacher Education and Special Education: The Journal of the Teacher Education Division of the Council for Exceptional Children* 31 (1): 36–46. doi:10.1177/088840640803100104.

McCormick, Michael J., Kim E. Dooley, James R. Lindner, and Richard L. Cummins. 2007. "Perceived Growth versus Actual Growth in Executive Leadership Competencies: An Application of the Stair-step Behaviorally Anchored Evaluation Approach." *Journal of Agricultural Education* 48 (2): 23–35.

Meier, Sherry L., Beverly S. Rich, and Jo Ann Cady. 2006. "Teachers' Use of Rubrics to Score Non-traditional Tasks: Factors Related to Discrepancies in Scoring." *Assessment in Education: Principles, Policy and Practice* 13 (1): 69–95.

Moni Roger, W., and B. Karen Moni. 2008. "Student Perceptions and Use of an Assessment Rubric for a Group Concept Map in Physiology." *Advances in Physiology Education* 32 (1): 47–54. doi:10.1152/advan.00030.2007.

Moni, Roger W., Eileen Beswick, and Karen B. Moni. 2005. "Using Student Feedback to Construct an Assessment Rubric for a Concept Map in Physiology." *Advances in Physiology Education* 29 (4): 197–203.

Mott, Michael S., Debby A. Chessin, William J. Sumrall, Angela S. Rutherford, and Virginia J. Moore. 2011. "Assessing Student Scientific Expression Using Media: the Media-enhanced Science Presentation Rubric (MESPR)." *Journal of STEM Education: Innovations and Research* 12 (1&2): 33–41.

Newman, Lori R., Beth A. Lown, Richard N. Jones, Anna Johansson, and Richard M. Schwartzstein. 2009. "Developing a Peer Assessment of Lecturing Instrument: Lessons Learned." *Journal of the Association of American Medical Colleges* 84 (8): 1104–1110.

Nicholson, Patricia, Shelley Gillis, and A. M. Dunning. 2009. "The Use of Scoring Rubrics to Determine Clinical Performance in the Operating Suite." *Nurse Education Today* 29 (1): 73–82.

Norcini, John J. 1999. "Standards and Reliability in Evaluation: When Rules of Thumb Don't Apply." *Academic Medicine* 74 (10): 1088–1090.

Pagano, Neil, Stephen A. Bernhardt, Dudley Reynolds, Mark Williams, and Matthew Kilian McCurrie. 2008. "An Inter-Institutional Model for College Writing Assessment." *College Composition and Communication* 60 (2): 285–320.

Panadero, Ernesto, and Anders Jonsson. 2013. "The Use of Scoring Rubrics for Formative Assessment Purposes Revisited: A Review." *Educational Research Review* 9: 129–144. doi:10.1016/j.edurev.2013.01.002.

Panadero, Ernesto, Jesús Alonso-Tapia, and Eloísa Reche. 2013. "Rubrics Vs. Self-assessment Scripts Effect on Self-regulation, Performance and Self-efficacy in Pre-service Teachers." *Studies in Educational Evaluation* 39: 125–132.

Panadero, Ernesto, Jesús Alonso Tapia, and Juan Antonio Huertas. 2012. "Rubrics and Self-assessment Scripts Effects on Self-regulation, Learning and Self-efficacy in Secondary Education." *Learning and Individual Differences* 22 (6): 806–813. doi:10.1016/j.lindif.2012.04.007.

Peach, Brian E., Arup Mukherjee, and Martin Hornyak. 2007. "Assessing Critical Thinking: A College's Journey and Lessons Learned." *Journal of Education for Business* 82 (6): 313–320.

Petkov, Doncho, and Olga Petkova. 2006. "Development of Scoring Rubrics for IS Projects as an Assessment Tool." *Issues in Informing Science and Information Technology* 3: 499–510.

Pintrich, Paul R., and Akane Zusho. 2002. "The Development of Academic Self-regulation: The Role of Cognitive and Motivational Factors." In *Development of Achievement Motivation*, edited by Allan Wigfield and Jacquelynne S. Eccles, 249–284. San Diego, CA: Academic Press.

Reddy, Malini Y. 2011. "Design and Development of Rubrics to Improve Assessment Outcomes: A Pilot Study in a Master's Level Business Program in India." *Quality Assurance in Education* 19 (1): 84–104.

Reddy, Y. Malini, and Heidi Andrade. 2010. "A Review of Rubric Use in Higher Education." *Assessment & Evaluation in Higher Education* 35 (4): 435–448. doi:10.1080/02602930902862859.

Reynolds-Keefer, Laura. 2010. "Rubric-referenced Assessment in Teacher Preparation: An Opportunity to Learn by Using." *Practical Assessment, Research & Evaluation* 15 (8) http://pareonline.net/pdf/v15n8.pdf.

Rezaei, Ali Reza, and Michael Lovorn. 2010. "Reliability and Validity of Rubrics for Assessment through Writing." *Assessing Writing* 15 (1): 18–39. doi:10.1016/j.asw.2010.01.003.

Reznitskaya, Alina, Li-jen Kuo, Monica Glina, and Richard C. Anderson. 2009. "Measuring Argumentative Reasoning: What's behind the Numbers?" *Learning and Individual Differences* 19 (2): 219–224. doi:10.1016/j.lindif.2008.11.001.

Rochford, Linda, and Patricia S. Borchert. 2011. "Assessing Higher Level Learning: Developing Rubrics for Case Analysis." *Journal of Education for Business* 86 (5): 258–265. doi:10.1080/08832323.2010.512319.

Sadler, Philip M., and Eddie Good. 2006. "The Impact of Self-and Peer-grading on Student Learning." *Educational Assessment* 11 (1): 1–31.

Schamber, Jon F., and Sandra L. Mahoney. 2006. "Assessing and Improving the Quality of Group Critical Thinking Exhibited in the Final Projects of Collaborative Learning Groups." *The Journal of General Education* 55 (2): 103–137.

Schlitz, Stephanie A., O. Connor Margaret, Yanhui Pang, Deborah Stryker, Stephen Markell, Ethan Krupp, Celina Byers, Sheila Dove Jones, and Alicia King Redfern. 2009. "Developing a Culture of Assessment through a Faculty Learning Community: A Case Study." *International Journal of Teaching and Learning in Higher Education* 21 (1): 133–147.

Schreiber, Lisa M., Gregory D. Paul, and Lisa R. Shibley. 2012. "The Development and Test of the Public Speaking Competence Rubric." *Communication Education* 61 (3): 205–233.

Siegel, Marcelle A., Kristy Halverson, Sharyn Freyermuth, and Catharine G. Clark. 2011. "Beyond Grading: A Series of Rubrics for Science Learning in High School Biology Courses." *Science Teacher* 78 (1): 28–33.

Spence, Lucy K. 2010. "Discerning Writing Assessment: Insights into an Analytical Rubric." *Language Arts* 87 (5): 337–352.

Stellmack, Mark A., Yasmine L. Konheim-Kalkstein, Julia E. Manor, Abigail R. Massey, and Julie Ann. P. Schmitz. 2009. "An Assessment of Reliability and Validity of a Rubric for Grading APA-Style Introductions." *Teaching of Psychology* 36 (2): 102–107. doi:10.1080/00986280902739776.

Stiggins, Richard J., and Nancy Faires Conklin. 1992. *In Teachers' Hands: Investigating the Practices of Classroom Assessment*. Albany: SUNY Press.

Timmerman, Briana E. Crotwell, Denise C. Strickland, Robert L. Johnson, and John R. Payne. 2011. "Development of a 'Universal' Rubric for Assessing Undergraduates' Scientific Reasoning Skills Using Scientific Writing." *Assessment & Evaluation in Higher Education* 36 (5): 509–547. doi:10.1080/02602930903540991.

Torrance, Harry. 2007. "Assessment as Learning? How the Use of Explicit Learning Objectives, Assessment Criteria and Feedback in Post-secondary Education and Training Can Come to Dominate Learning." *Assessment in Education: Principles, Policy & Practice* 14 (3): 281–294.

Vandenberg, Amy, Matthew Stollak, Linda Mckeag, and Doug Obermann. 2010. "GPS in the Classroom: Using Rubrics to Increase Student Achievement." *Research in Higher Education Journal* 9: 1–10.

Wald, Hedy S., Jeffrey M. Borkan, Julie Scott Taylor, David Anthony, and Shmuel P. Reis. 2012. "Fostering and Evaluating Reflective Capacity in Medical Education: Developing the REFLECT Rubric for Assessing Reflective Writing." *Academic Medicine: Journal of the Association of American Medical Colleges* 87 (1): 41–50. doi:10.1097/ACM.0b013e31823b55fa.

Wallace, Colin S., Edward E. Prather, and Douglas K. Duncan. 2011. "A Study of General Education Astronomy Students' Understandings of Cosmology. Part II. Evaluating Four Conceptual Cosmology Surveys: A Classical Test Theory Approach." *Astronomy Education Review* 10: 010107. doi:10.3847/AER2011030.

Yopp, By David, and Richard Rehberger. 2009. "A Curriculum Focus Intervention's Effects on Prealgebra Achievement'." *Journal of Developmental Education* 33 (2): 28–30.

Zimmerman, Barry J. 2011. "Motivational Sources and Outcomes of Self-regulated Learning." In *Handbook of Self-regulation of Learning and Performance*, edited by Barry J. Zimmerman and Dale H. Schunk, 49–64. New York: Routledge.