A CRITICAL REVIEW OF THE ARGUMENTS AGAINST THE USE OF RUBRICS

Ernesto Panadero [1 2] & Anders Jonsson [3]

Author Note

[1] Facultad de Psicología y Educación. Universidad de Deusto, Spain.

[2] IKERBASQUE, Basque Foundation for Science, Bilbao, Spain.

[3] Faculty of Teacher Education. Kristianstad University, Sweden.

Recommended citation:
Panadero, E. & Jonsson, A. (2020). A critical review of the arguments against the use of rubrics. Educational Research Review (*online first*). doi: 10.1016/j.edurev.2020.100329

Correspondence concerning this article should be addressed to: Ernesto Panadero. Despachos IRPO. Universidad Deusto, 48007, Bilbao. Spain. E-mail: ernesto.research@gmail.com

Acknowledgements: The authors would like to thank David Boud for feedback on an earlier version of this manuscript; to the three rubric experts that were consulted during the literature search: Susan Brookhart, Heidi Andrade and Philip Dawson; to the three journal reviewers, especially Reviewer 2 that went beyond the line of duty; and to the four experts that provided feedback on the final proofs: Susan Brookhart, Anastasiya Lipnevich, Royce Sadler and Dylan Wiliam.

Research funded by: personal grant to first author under Ramón y Cajal framework (RYC- 2013-13469); Spanish Ministry of Economy and Competitiveness (Ministerio de Economía y Competitividad) National I+D Call (Convocatoria Excelencia) project reference EDU2016-79714-P; and Fundación BBVA call Investigadores y Creadores Culturales 2015 (project Transición a la educación superior id. 122500).

Abstract

Rubrics are widely used in classrooms at all educational levels across the globe, for both summative and formative purposes. Although the empirical support for the benefits of using rubrics has been steadily growing, so have the criticisms. The aim of this review is to explore the concerns and limitations of using rubrics as proposed by the critics, as well as the empirical evidence for their claims. Criticisms are then contrasted with findings from studies reporting empirical evidence in the opposite direction (i.e. supporting the use of rubrics). A total of 27 publications were identified, and 93 excerpts were extracted, after a detailed content analysis. The criticisms were organized around six broad themes. One of the main findings is that the empirical evidence behind criticisms is, with only a few exceptions, neither direct nor strong. On the contrary, several critics refer to anecdotal evidence and/or personal experiences, which have limited value as scientific evidence. Another finding is that a number of critics make claims about rubrics with a narrow conceptualization of rubrics in mind. One prevalent assumption is that rubrics are only used for high stakes testing and/or other summative assessment situations. Based on these findings, we advocate a more pragmatic approach to rubrics, where potential limitations of rubrics are investigated empirically and decisions are based on scientific data.

Keywords: Rubrics; Formative assessment; Criteria compliance; Explicit criteria; Self-regulated learning.

Highlights

- Rubrics are popular assessment tools that have received considerable criticism
- Our aim was to explore the focus and empirical evidence behind this criticism
- 27 publications were identified, 93 excerpts were extracted and organized into six themes
- Only a few studies provided empirical support for their claims
- Rubrics seem to have more benefits than drawbacks, especially when used formatively

A CRITICAL REVIEW OF THE ARGUMENTS AGAINST THE USE OF RUBRICS

Rubrics and the use of rubrics is currently one of the "hottest" research topics in education and educational psychology. There are two observations supporting this claim. First, the number of publications exploring the effects of the use of rubrics has been growing almost exponentially during the last ten years (Dawson, 2017). Second, within this relatively short time span, at least four empirical reviews about rubrics have been published in impact factor journals (Brookhart & Chen, 2015; Jonsson & Svingby, 2007; Panadero & Jonsson, 2013; Reddy & Andrade, 2010). From this bulk of empirical research, we know that the use of rubrics can have significant and positive effects on students' learning, academic performance, and self-regulation – provided that the design and implementation are adequate (Brookhart & Chen, 2015; Panadero & Jonsson, 2013).

Still, there are many publications criticizing the use of rubrics, ranging from the general idea of rubrics (e.g. Kohn, 2006), to rubrics being ill-designed and/or poorly implemented, and therefore being more harmful than beneficial (e.g. Wilson, 2007). In addition, there is literature arguing that sharing explicit criteria with students is inadequate and almost inevitably leads to instrumental learning and "criteria compliance" among students (e.g. Sadler, 2014; Torrance, 2007), or have other negative effects on teachers, teaching, or students. The latter view has obvious repercussions for the use of rubrics, since rubrics are probably the most common way of sharing explicit assessment criteria with students. All these concerns are widespread, despite the accumulated empirical support for mainly positive effects.

The aim of this review is therefore to explore the claims made by the critics, in order to identify the main areas of criticism and investigate to what extent these claims

are supported by empirical research. This means that the intent is not to dismiss or prove the critics wrong, but to explore the grounds for their claims and compare the empirical support, to the accumulated empirical research, as presented for instance in the systematic reviews referred to above.

Definition, design, and purposes

Rubrics are instruments designed for assisting assessors in judging the quality of student performance and/or help teachers and learners to judge the quality and progression of student performance. Rubrics have three fundamental features in common (Popham, 1997). First, to assist in identifying the qualities to be assessed, the rubric includes information about which aspects or criteria to look for in student performance. If the rubric is analytic, the assessment criteria are typically presented in separate rows, while if it is holistic the criteria are integrated in the descriptions of the performance levels. Second, to assist in judging the quality of student performance, the rubric includes descriptions of student performance at different levels of quality. By combining the aspects to be assessed with the descriptions of quality into a two-dimensional matrix, along with the third feature, which is a “scoring strategy”, a rubric comes into existence. An example of a rubric is shown in Figure 1.

	4	3	2	1
Contributions	In large- and small-group discussions, shares helpful ideas. Leads the discussion and makes a strong effort to contribute.	In large- and small-group discussions, often shares helpful ideas. Clearly strives to participate.	In large- and small-group discussions, sometimes shares helpful ideas. Makes the required effort to participate but no more.	In large- and small-group discussions, rarely shares helpful ideas. Participates minimally or not at all.
Time management	Completes assignments on time throughout the project. Does not cause the group to change deadlines or reassign work because of lateness.	Usually completes assignments on time throughout the project. Does not cause the group to change deadlines or reassign work because of lateness.	May put things off, but turns assignments in on time. Does not cause the group to change deadlines or reassign work because of lateness.	Routinely misses deadlines, puts off work, and causes group to change deadlines or reassign work because of lateness.
Problem solving	Makes a clear effort to find and share answers to problems.	Does not actively seek answers to problems but helps to improve those found by others.	Accepts solutions found by others without changing them. Is willing to try suggested answers to problems.	Makes no effort to find, share, or try answers to problems. Leaves all work to others.
Working with others	Listens well and assists others in their efforts. Facilitates group work.	Usually listens well and assists others in their efforts. Does not facilitate group work, but doesn't hinder it either.	Sometimes listens well and assists others in their efforts but may be difficult to work with.	Does not listen well or assist others; may not participate in group work.
Research techniques	Always looks at varied sources and records information in detail.	Usually studies varied sources and records information in some detail.	Often studies varied sources and records information, but sometimes it is sketchy.	Rarely looks at more than one source and barely takes any notes.
Synthesis	Arranges information found by self and others into useful formulations; is able to manage complex ideas.	Usually arranges information found by self and others into useful formulations; may need help in managing complex ideas.	Sometimes arranges information found by self and others into useful formulations. Does not manage complex ideas.	Rarely or never arranges information into useful formulations or manages complex ideas.

Figure 1. An example of a widely used analytic rubric: the collaboration rubric from the International Reading Association and the National Council of Teachers of English.

Rubrics can differ in design and implementation depending on a number of different factors, most notably how the assessment information should be used (e.g. for summative or formative purposes). For example, rubrics can be designed as either holistic or analytic (Brookhart, 2013). A holistic rubric (also known as “comprehensive” or “global”) is arranged around a single scale, thereby treating quality as unidimensional, with all criteria to be included in the evaluation being considered together. Therefore, using this type of rubric, assessors assign an overall judgment of student work. In an analytic rubric, on the other hand, all criteria are presented separately, facilitating a more detailed and multidimensional assessment. Holistic rubrics are therefore typically used for summative assessments, where detail is not always needed, while analytic rubrics are more suitable for formative assessments (Brown, 2018). By presenting the criteria separately, strengths and weaknesses in student performance can be easily identified and this information can be used by students to improve future task performance.

Another example is the “specificity” of rubrics. In terms of content, rubrics can be more or less general or specific (Brookhart, 2013). In a general rubric, the descriptions of student performance are based in characteristics of a general competence or task (e.g. writing). A task-specific rubric, on the other hand, is designed to support the assessment of one particular task. Such a rubric may be quite detailed and include correct/appropriate student responses, which would give away the answers if shared with students. There are also intermediate “task-type rubrics” (Dawson, 2017), which may be used to support the assessment of a group of similar tasks, such as argumentative texts or oral presentations. This means that raters and/or students may use the same rubric at several occasions and over time internalize the criteria.

Another decision that has to be made when designing a rubric is the number of criteria and performance levels that the rubric should include. Here, the demand for interrater agreement in summative assessments often makes it appropriate to have fewer levels (Jonsson & Svingby, 2007), although high-stakes assessments may require longer scales in order to achieve adequate reliability. In formative assessments, there is a need for fine-grained information about student performance, but also a higher tolerance for error (since the assessment can be changed if it turns out to be erroneous), making several performance levels useful in order to identify specific strengths and weaknesses.

Although there is no such thing as the ideal rubric, and although there is still much work to be done in systematically investigating different aspects of design and use of rubrics, the examples above suggest that rubrics can be designed and implemented in ways that are more or less in line with the intended use. Based on the reviews by Jonsson and Panadero (2016) and Brookhart (2018), some general recommendations for the design and implementation of rubrics for formative purposes could be to:

- Use an analytic scoring strategy without summarizing into a total score, so that the aspects to be assessed are explicitly spelled out and strengths and weaknesses in relation to individual criteria are discernible.
- Use multiple quality levels, so that the quality sought becomes visible to the students and for aiding in producing and understanding future-oriented feedback.
- Use task-level specificity, so that rubrics are neither too closely tied to the particular task nor too general. Instead, rubrics need to be applicable to several, but similar, tasks assessing the same performance.
- Support students in internalizing the criteria by (a) explaining the criteria and quality levels, (b) making the rubric available, digitally or on paper, (c)

providing the students with the rubric before they perform the task, and (d) model the use of rubrics.

- Use descriptive language, rather than evaluative language, such as “excellent” or “poor” (Brookhart, 2018), so that the descriptions may guide student performance and facilitate self-assessment and reflection.

As a more advanced implementation, when students have gained some expertise in the area of interest, they could be invited co-create the rubric with the teacher.

Effects on student performance from using rubrics

In 2007, Jonsson and Svingby (2007) published a review on the use of rubrics for both summative and formative purposes. Findings from their review suggest that the use of rubrics can promote student learning and/or improving instruction by making expectations and criteria explicit, which was seen to facilitate assessment processes such as feedback and self-assessment.

It is important to note that the number of studies investigating the formative potential of rubrics was quite limited in 2007 (the review by Jonsson and Svingby included only 25 studies). Since then, the interest in rubrics has grown steadily. Dawson (2017) writes that the 100th paper mentioning “assessment rubrics” was published in 1997, the 1000th in 2005, and sometime in 2013, the 5000th paper mentioning rubrics was published. Furthermore, in 2013, a new review on rubrics was published, which focused exclusively on the formative function of rubrics (Panadero & Jonsson, 2013). The findings from that review corroborated previous findings in that the use of rubrics may provide transparency to the assessment, which in turn may: (a) reduce student anxiety, (b) aid the feedback process, and (c) support student self-regulation; all of which may indirectly facilitate improved student performance. Brookhart and Chen (2015) also noted, in a follow-up review on both summative and formative uses of

rubrics, that several studies reporting on the effects of rubric use on learning and performance used relatively rigorous designs, such as experiments and quasi-experimental studies. Since then, the number of empirical studies showing that rubrics may have a positive effect on student performance has continued to grow (e.g. Greenberg, 2015; Lipnevich, McCallen, Miles, & Smith, 2014).

How rubrics support student learning

As outlined above, rubrics may support student learning in different ways, such as facilitating the understanding of expectations and feedback, as well as through supporting students' self-regulated learning (SRL). Panadero and Jonsson (2013) suggested that rubrics provide a sense of transparency, which is generally appreciated by students. For example, Andrade and Du (2005) reported that pre-service teachers claimed to use rubrics in order to determine their own teacher educators' expectations. The students also compared the frustration of not knowing their teacher's expectations with the relief provided by a rubric. Importantly, some students talked about using rubrics to orient themselves toward their teacher's expectations, while still allowing them to "make the decisions ourselves about how we wanted to go about it" (p. 4). Similar findings are reported by Reynolds-Keefer (2010), who used questionnaires to investigate the perceptions of using rubrics among students in educational psychology. Of course, students' appreciation of rubrics should not be confused with their actual understanding of the criteria. However, as shown by Holmstedt, Jonsson, and Aspelin (2018), by providing explicit criteria to pre-service teachers, the students were able to discern and discuss important aspects of teaching not previously noted.

Studies have reported students using rubrics to plan, monitor, and evaluate their task performance (i.e., SRL). For example, Jonsson (2014) performed case studies in different settings in professional education (public health, real estate brokers, and dental

education), where rubrics were used. In these settings, students claimed to use the rubric as guidance when planning their assignments. Several students also reported that they used the rubrics in order to assess the progress of their work during task performance, using the criteria "as targets to make sure we included all the important aspects in the assignment" (p. 10), as well as to make a final check before submitting the assignment for summative assessment. Similar findings are reported in other studies, including the studies by Andrade and Du (2005) and Reynolds-Keefer (2010) mentioned above.

Panadero and his colleagues investigated dimensions of SRL associated with negative emotions and stress, actions directed by anxiety, external pressure to perform, and task avoidance. In these studies, students' scores on performance- and avoidance-oriented SRL scales were shown to decrease for students using rubrics (Panadero et al., 2013; Panadero & Romero, 2014). These are positive findings, but the students using rubrics in the study by Panadero and Romero (2014) also reported higher levels of stress while performing the task as compared to the control group. Furthermore, learning-oriented SRL scores may decrease for students using rubrics (Panadero, Alonso-Tapia, & Huertas, 2014). This means that while the use of rubrics may decrease performance- and avoidance-oriented SRL strategies, typically unfavourable for learning, they do not necessarily increase the use of learning-oriented SRL strategies.

Another aspect of supporting students' SRL, is students' self-efficacy, which may affect future task performance. By using rubrics to improve performance and facilitate self-assessment, rubrics should (ideally) increase student self-efficacy and favour a mastery-orientation among the students. This ideal picture, however, does not really appear in empirical research. For example, in a study by Panadero, Alonso-Tapia, and Reche (2013) no significant effects were found for students' self-efficacy. Furthermore, in a meta-analysis on the effects of self-assessment interventions on self-

regulation strategies, students using rubrics (or similar instruments), reported lower self-efficacy after the intervention than participants not using them (Panadero, Jonsson, & Botella, 2017). This could be an effect of students becoming aware of the complexity of high-quality performance and therefore reporting lower self-efficacy. These findings correspond to Andrade and Du (2005), where some students found it overwhelming to read all the levels of the rubric before receiving feedback from the teacher. However, there were also students aiming for higher performance, who did not read the levels describing lower quality in the rubric, suggesting that students may react differently to rubrics, depending on their previous performance and academic self-esteem.

Criticism on the use of explicit criteria and rubrics

In addition to research suggesting that rubrics may facilitate student learning, there are critics arguing against the use of explicit criteria for formative purposes. In some cases, the opposition is only a matter of perspective, but in others, it is not. Analytic assessments, for example, focus on the parts rather than the whole, which may involve a risk of fragmentation; Sadler (2009) therefore argues against the use of analytical assessment and pre-set criteria, in favour of holistic assessment with "emergent" criteria. Emergent criteria mean that assessors should not set any criteria beforehand, but address criteria that surface in the moment of assessing a particular piece of work. One of Sadler's main arguments for this approach is what he refers to as the "indeterminacy of criteria": When breaking down holistic judgments into more or less discrete components, these components – no matter how many they are and no matter how carefully they are selected – cannot sufficiently represent the full complexity of the multi-criterion qualitative judgment made by the assessor (cf. Popham, 1994). To substantiate this argument, Sadler made observations on how assessors approach assessment and/or grading. Most referred to differences between

holistic and analytic judgements, such as assessors agreeing on the overall grade/score for a particular work, but not on the level of performance for individual criteria. Sadler also noted that, in his experience, teachers generally have more confidence in their own holistic judgements as compared to analytical assessments and that global judgments are often made through the lenses of the pre-set criteria. The latter means that qualities not visible through those lenses might be filtered out and not taken into account.

There are also arguments against sharing pre-set criteria and rubrics with students. For example, Sadler (2014) argues that students may not understand the criteria, since words, symbols, diagrams, and other "codifications" lack the necessary attributes to represent the criteria or standards. Any attempt to communicate academic expectations through such means is therefore "fundamentally flawed" and Sadler argues that students need to develop a conceptualization of what constitutes "quality" by continuously evaluating authentic work, without being hampered by criteria specified beforehand.

Torrance (2007; 2012) also claims that the use of explicit criteria may turn students' attention away from productive learning and focus on surface strategies and "criteria compliance" instead. According to Torrance, the core aspiration of (higher) education should be on students' autonomous thinking, rather than on the convergent thinking produced by transparency in assessment processes and criteria.

These criticisms against the use of pre-set criteria are only examples from a more extensive literature on the topic. Still, they are interesting because they make strong claims about the limitations and/or consequences from using explicit criteria for formative purposes, which are based on either personal or theoretical considerations, or empirical conditions that are difficult to generalize. This raises interesting questions

about the basis for the criticism against the use of explicit criteria and rubrics, such as to what extent these claims are supported by empirical research.

Aim of the study

Taken together, current research suggests that rubrics, when properly designed and implemented, may support student performance, for example by facilitating the understanding of expectations and supporting students' SRL. However, despite this empirical evidence, there is still a number of critics, opposing to the use of rubrics for formative purposes. The aim of this review is therefore to scrutinize the claims made by the critics, in order to investigate: (a) What are the main areas of criticism, (b) to what extent are these claims supported by empirical research, and (c) whether there is empirical research presenting evidence against these claims?

Method

Selection of studies

Three search methods were used in the initial phase. First, we conducted independent searches in PsycINFO, ERIC, and Google Scholar. Both authors used the following combinations of keywords: Rubric/s + critique/s, Rubric/s + flaw/s; and the first author searched at the latter phase using: Rubric/s + reliab* (for reliability), and Rubrics/s + valid* (for validity). This strategy offered a low number of relevant hits and only 8 publications were selected for further exploration. Second, as both authors have been following the topic of rubrics for more than a decade, we combined our reference libraries summing up to 280 references to identify potential publications to include (see appendix 1 for the list). This strategy offered the largest proportion of hits, resulting in 24 publications for further exploration. Third, we consulted three experts on research about rubrics (see acknowledgements) for publications they thought should be considered for this review. Their recommendations overlapped with the ones taken from

our own libraries, adding only one additional publication. After a full article read of these 33 publications, 20 were included in the final review. In the final phase of the search, using the "snowball method" based on citations or quotes from already included articles, another 29 publications of interest were identified, 10 of which had already been identified in the previous phases. After reading the new publications, 7 were included. The final number of publications included was 27; 26 journal articles and one book. Most publications that were not included was on the grounds of not presenting any criticism to rubrics, though initially it seemed like they were.

As an attempt to include as many claims against rubrics as possible, peer-review or empirical grounding were *not* used as inclusion criteria for the publications. This approach was also used because analysing the quality of the empirical evidence behind each claim is part of the aim of the review. The three inclusion criteria were that the publications should: (a) cover criticism of rubrics, (b) be published in journals or books (regardless of review process), and (c) be written in English.

Extracting information from the articles and coding

The analysis was initiated by creating a database with verbatim excerpts from the original articles, where critical remarks about rubrics were made. These excerpts (n=93) were used as data for the analysis (see Appendix 2). Once the database was finalized, different themes were identified in the data. First, each author independently analysed each excerpt, extracting the main idea behind the criticism, and created an initial categorization of all excerpts. When the categories from both authors were compared, there was agreement for 69 of the excerpts (74%), and the rest was discussed case by case, aiming for consensus. Once total agreement was reached, each author independently merged the initial categories into more general themes. When compared, the level of agreement was 94%. Three excerpts which have different themes were

grouped in a new miscellanea category. Additionally, one author had assigned two themes to three excerpts. For two of them, a single theme was decided, making the final number of themes six. For the other excerpt it was decided to keep the double classification, due to the extensive nature of the excerpt (Wilson, 2006).

Data analysis

This review is of narrative type, using, first, content analysis of the original sources to identify both the categories of criticism to rubrics and their content, and, second, thematic analysis in order to group the data into more general themes. Furthermore, all excerpts were analysed to identify the empirical evidence behind each claim in two ways. First, the excerpt itself was analysed in order to see whether there was scientific evidence to support the claims made. Second, when the excerpt included references to other publications, these were explored for scientific evidence to support the original claim. This was performed up to a second level (i.e. if that new reference also includes new relevant references, these were also analysed for scientific evidence). A third level was never needed.

Since the analysis involves the categorization of "scientific evidence", a distinction has to be made regarding what is considered "scientific" or not. Although the definitions of "science" and "research" may differ, "systematic inquiry" seems to be a common denominator. Here, "scientific evidence" therefore refers to the outcome of any kind of activity, which is characterized by "systematic inquiry". This includes, for example, RCT and quasi-experimental designs, as well as case studies and ethnographic investigation, regardless of whether the data is quantitative, qualitative, or mixed. "Scientific evidence" may also include self-observations and autobiographies, as long as the inquiry process is reported and made transparent. It does not, however, include anecdotal data, based on the author's own experiences with rubrics or hearsay, since

such data is not systematically collected or analysed. Anecdotal evidence may therefore be strongly biased. While this is a risk to most qualitative research designs, the demands for transparency when reporting on methodology may counteract validity issues and potential bias. When referring to personal experiences, without providing any methodological information, this has not been categorized as “scientific evidence”. A second distinction made, is between “empirical” and “theoretical/conceptual” evidence. Both are considered scientific, by systematically investigating either empirical data or theories, models, and concepts, but they are also considered *reciprocally interdependent*. This means that in order to interpret and understand empirical data, theoretical contributions are needed, for instance, to make important conceptual demarcations or to provide different perspectives. However, it also means that theories need to be empirically validated. If theories are not empirically validated, they may mislead us, for instance by predicting events that are extremely unlikely to occur (even if theoretically possible). This risk is particularly imminent in relation to educational theories, since the “real-world context”, to which the theories relate, is so complex and difficult to predict. As a consequence, in this review a distinction is made between, on the one hand, theoretical evidence that has *not* been empirically validated and, on the other, empirical evidence and/or theoretical evidence that has been empirically validated.

Results

This section is organized around the themes that were identified. Table 1 shows the identified categories of criticisms, the publications and the specific excerpts numbers. In addition, Appendix 2, which is provided as an online supplementary material, includes a table with the verbatim quotes and the empirical evidence behind each claim.

Table 1. Listing of categories and excerpts

Main category	**Put forth by (with numbering)**
Standardization and narrow the curriculum	Popham, 1997 (1, 2, 3, 4)
	Mabry, 1999 (5, 6, 10, 11, 12, 15)
	Kohn, 2006 (25, 27, 28, 29, 32)
	Wilson, 2006 (33)
	Wilson, 2007 (35, 36)
	Chapman & Inman, 2009 (37)
	Sadler, 2009 (46)
	Rezaei & Lovorn, 2010 (55)
	Lovorn & Rezaei, 2011 (60)
	Shipman, Roa, Hooten & Wang, 2012 (65)
	Humphry & Heldsinger, 2014 (72)
	Ito, 2015 (78)
	Bouwer, Lesterhuis, Bonne & De Maeyer, 2018 (92)
Validity	Mabry, 1999 (7, 8, 9, 14)
	Andrade, 2005 (21)
	Sadler, 2009 (41, 47)
	Humphry & Heldsinger, 2014 (68, 69, 70)
Instrumentalism and "criteria compliance"	Mabry, 1999 (13)
	Norton, 2004 (16, 17)
	Andrade, 2006 (22, 23)
	Kohn, 2006 (30, 31)
	Wilson, 2006 (33)
	Torrance, 2007 (34)
	Chapman & Inman, 2009 (38)
	Torrance, 2012 (66)
	Bell, Mladenovic & Price, 2013 (67)
	Bearman & Ajjawi, 2018 (90, 91)
Simple implementations do not work	O'Donovan, Price & Rust, 2004 (18)
	Andrade, 2005 (19, 20)
	Rezaei & Lovorn, 2010 (53, 54)
	Lovorn & Rezaei, 2011 (58, 59, 61)
	Jones et al., 2016 (81, 82, 83)
	Wollenschläger, Hattie, Machts, Möller & Harms, 2016 (84)
Limitations of criteria	Kohn, 2006 (26)
	Sadler, 2009 (39, 40, 42, 43, 44, 45, 51, 52)
	Bloxham, Boyd & Orr, 2011 (56, 57)
	Shipman, Roa, Hooten & Wang, 2012 (62, 63, 64)
	Sadler, 2014 (73)
	Ito, 2015 (74, 75, 79)
	Jones et al., 2016 (80)
	Hudson, Bloxham den Outer & Price, 2017 (85)
	Bearman & Ajjawi, 2018 (86, 88)
Context dependence	Sadler, 2009 (48, 49, 50)
	Humphry & Heldsinger, 2014 (71)
	Ito, 2015 (76)
	Bearman & Ajjawi, 2018 (87)
	Brookhart, 2018 (93)
Miscellanea	Kohn, 2006 (24)
	Ito, 2015 (77)
	Bearman & Ajjawi, 2018 (89)

Note: the content and the empirical evidence for each excerpt can be found in Appendix 2 published online as supplemental material

Theme #1: Standardization and narrowing the curriculum

This theme covers criticisms on how rubrics standardize assessments by providing simple lists of criteria for complex skills and creating a tendency for students and teachers to guide their actions toward those criteria. Per these authors, the criteria narrow the curriculum because they restrict performance of complex skills. The majority of cases refer to writing and the assessment of writing performance.

Summary of main criticisms

Three subthemes were identified: a) Standardization of assessment through rubrics, b) Rubrics narrowing the curriculum, and c) The reduction in variability of scores. According to some of the excerpts, standardization produces a narrow assessment and conceptualization of the task. Excerpt 37 is an example: "We were struck by a child's veracity about the restrictions a rubric-oriented teaching force places on our learners. Such restrictions may be real: the students must adhere strictly to prescribed criteria with no deviations, per the teacher's instructions, or student culture may impose restrictions" (Chapman & Inman, 2009, p. 53). As a consequence, there is a strong relationship between the two subthemes a) and b).

a) Standardization of assessment via rubrics (excerpts 2, 4, 5, 6, 11, 12, 15, 21, 25, 27, 29, 35, 37, 65). According to the excerpts, rubrics are carriers of a large-scale assessment and high-stakes accountability tests rationales, where quantitative comparisons, a strive for higher interrater agreement, and standardization are integral components. It is interesting to notice that this particular subtheme is mostly evident in earlier publications and, since 2009, the argument has turned to how standardization narrows the learning environment. The most extreme position in relation to this subtheme is taken by Wilson (2006), who resents any kind of standardization, and suggests that assessments

should be based on the immediate reactions evoked by student performance: “I suggest that we make ourselves transparent as we read – that we pay attention to what goes on in our minds and try to put our reactions and questions and wonderings and musings and connections and images into words” (p. 63).

b) Rubrics narrow the curriculum. The general concern is the impact rubrics have as biased representations of complex skills and how they influence instruction. This criticism can be found in earlier publications (excerpts 1, 3, 9, 10, 32, 33) and then more directly connected to standardization in more recent publications (excerpts 37, 46, 55, 60, 65, 72, 78, 92). The excerpts cover a range of possible aspects that can be narrowed: students understanding and performance of the task (37), the criteria and standards teachers use for assessment (72), or the quality of peer feedback (92).

c) Less variability of scores (excerpts 8, 68, 69, 70). Mabry (1999) claims that the consistency in assessment provided by rubrics is mainly achieved because rubrics limit the scope of variability of scores. Humphry and Heldsinger (2014) also claim that rubrics, where all the different criteria have the same number of performance levels, produce a kind of halo effect. According to the authors, this effect can be explained by the fact that equal numbers of performance levels may result in either more or fewer levels for any given criterion, as compared to the number of distinctions that judges could potentially make. If there are more levels than judges can distinguish, they will have to make a global judgment instead. Alternatively, if there are fewer levels, judges cannot make the distinctions they are actually capable of. The halo effect could also be explained by unintended conceptual overlap and redundancy in the criteria and/or performance levels. This means that, in order to avoid such halo effects, rubrics

> should preferably not have the same number of performance levels for all criteria.

Empirical evidence backing up the claims

Of all the studies that covered this particular theme, only two present direct empirical evidence to support the claims made.

First, Bouwer et al. (2018) investigated two different instructional approaches: applying criteria to examples and comparative judgment. Students were instructed to write a short essay, preceded by a peer-assessment session in which they evaluated the quality of a range of example essays. Half of the students (n=20) evaluated the quality of the example essays using a list of teacher-designed criteria, while the other group (n=20) evaluated the essays by pairwise comparisons. Results show that the students in the comparative judgment condition provided relatively more feedback on higher order aspects, as compared to students in the criteria condition. This was only the case for improvement feedback, however, while there were no significant differences for feedback on strengths. The findings from this study therefore suggest that the enacted curriculum is indeed narrowed by the rubric condition, as least as compared to the pairwise comparisons.

The second study, by Humphry and Heldsinger (2014), used an "iterative tryout-redesign-tryout approach" (p. 253) to investigate a rubric used to assess students' writing in a large-scale testing program. These authors present empirical evidence suggesting that it is problematic for rubrics to have the same number of performance levels for all criteria when:

> /…/ there is no underlying developmental or learning theory that justifies having precisely the same number of qualitatively distinguishable stages across multiple aspects of a construct. (p. 253)

This is because it is:

> /…/ unlikely that the gradations of quality faithfully capture that which is observed in student performances for each criterion separately from other criteria. It is more likely instead that the same numbers of gradations of quality are chosen for convenience in constructing rubrics and for ease of marking than because equal numbers of gradations faithfully capture the distinguishable performance levels for separate criteria (p. 253)

Regarding the rest of the sources, the majority present anecdotal evidence, although at several occasions this is not made explicit. Furthermore, in a couple of cases, we have identified doubtful interpretations of own results, due to methodological inadequacies and an incorrect use of references (Lovorn & Rezaei, 2011; Rezaei & Lovorn, 2010). Mabry (1999), who is an expert on assessment with ample experience, makes a special case as she refers to a number of sources, several of which are personal communications and non-published material, as well as anecdotal evidence. Even though she is critiquing rubrics generally, her criticisms seem to be aiming for rubrics used for accountability testing purposes, which were widespread in the 1990s in the USA, and not about the use of rubrics for formative assessment purposes. According to Mabry (1999), there was a strong tendency for these large-scale testing rubrics to "give priority to the mechanical, format, and organizational aspects of writing" (p. 676). Some of her examples also suggest that some of these "rubrics" were actually rating scales, rather than rubrics. Still, the main claim is that the *standardization* of a self-expressive and individualistic skill, jeopardizes student learning and understanding of writing.

Empirical evidence against the claims

According to reviews of empirical research on rubrics, the use of rubrics may improve students' performance (Brookhart & Chen, 2015; Panadero & Jonsson, 2013).

While some of the critics would argue that it is because rubrics simplify what is to be learned, this would first and foremost depend on the validity of the rubric used. If the rubric aligns with what Andrade (2005) refers to as "respectable standards and with the curriculum being taught" (p. 29), high ratings according to the rubric should presumably coincide with teachers' judgments. If not, this would be substantial evidence of rubrics narrowing the curriculum. However, without providing information about the validity of the rubric, it is difficult to make any claims about to what extent rubrics narrow or simplify what is to be learned. Rubrics giving priority to the mechanical, format, and organizational aspects of writing (Mabry, 1999), may not be considered valid. But the question becomes more difficult to answer if considering rubrics based on, for instance, the "6+1 Trait Writing Model", which include a more comprehensive set of writing criteria, and/or rubrics that can be used more flexible as part of classroom formative assessment, rather than being part of a large-scale testing program. Should such, potentially more valid, rubrics also be considered to narrow the curriculum?

According to the reviews by Jonsson and Svingby (2007), and Brookhart and Chen (2015), there are studies presenting empirical evidence of, for example, content-, and criterion-related validity of rubrics. The main limitation of this research, as pointed out by Jonsson and Svingby (2007), is that most studies focus on only one or two aspects of validity. For instance, Brookhart and Chen (2015) wrote that while studies of the effects of the use of rubrics on student learning and motivation provide consequential evidence for the validity of formative use of rubrics, other aspects of validity are also needed. Both reviews therefore suggest that the validation of rubrics could be facilitated by using a more comprehensive framework of validity, such as Messick (1996) or Kane, Crooks, and Cohen (1999). Per Brookhart and Chen (2015), the validity evidence summarized in their review is impressive in some respects, but

they also emphasize that there is still more work to be done in relation to validity. The study by Humphry and Heldsinger (2014) can be considered as contributing to this call for additional validity evidence.

A narrow understanding of criteria and assessment might also produce other negative effects, such as standardization of learning goals and "teaching to the test". However, this is an issue that goes beyond the use of rubrics and refers to the classical tension between high-stakes and low-stakes assessments. In the US, for example, there has been a strong emphasis on high stakes testing accountability systems, usually following state-wide standards and external evaluation, which has affected classroom instruction (Au, 2007; Minarechová, 2012; Nichols & Harris, 2016). However, rubrics can be used both state-wide and within the individual classroom, but for different purposes. In relation to high-stakes assessments, rubrics can be used to support a summative use of rubrics, through standardization and increased interrater agreement, which may involve a reduction of the number of criteria used, and a focus on more easy-to-agree-on criteria. For formative uses, there is less need for standardization and some teachers may even have the freedom to create their own rubrics or creating them with colleagues or students. For formative purposes, the rubrics can also be used more flexibly, for instance as a scaffold that is gradually faded when students have internalized the criteria.

Theme #2: Instrumentalism and "criteria compliance"

A common assertion is that sharing assessment criteria with learners leads to instrumentalism and surface learning strategies, which has been termed "criteria compliance" by Torrance (2007). In essence, this means that when being informed about the criteria, learners will focus on meeting these criteria with minimal effort and

also limit their performance to what is explicated by the criteria, leaving other things aside (i.e. a special case of narrowing the curriculum, as outlined above).

Summary of main criticisms

Three main subthemes can be identified:

a) The steering effect of assessment criteria. This is the most dominant theme in the literature reviewed. For instance, Bearman and Ajjawi (2018) claim that "Many students seek to use the written criteria to pass the assessment rather than learn" (p. 4), suggesting that learners will try to meet the criteria in a superficial manner. The steering effect, which is often referred to as the "backwash" of assessment (Biggs, 1999), is clearly seen in the arguments of some critics (excerpts 13, 16, 17, 22, 23, 30, 31, 33, 38, 66, 67, 90). Kohn (2006), for example, quotes a sixth grader, who says that: "The whole time I'm writing, I'm not thinking about what I'm saying or how I'm saying it. I'm worried about what grade the teacher will give me, even if she's handed out a rubric. I'm more focused on being correct than on being honest in my writing." (p. 14).
b) Extensive use of coaching (excerpt 34). Torrance (2007) connects the transparency provided by criteria to the extensive use of coaching by teachers, tutors, and others. This coaching practice is argued to remove the challenges of learning for the students and to reduce the quality of the outcome. Basically, an overly supportive pedagogy, including the use of transparency in assessment, is argued to make the students passive and replace "learning" with conformity and "criteria compliance".
c) Transfer of "epistemic beliefs" (excerpt 91). In all pedagogical situations, educators implicitly send messages about how to understand knowledge itself. This is especially true for assessment situations, which clearly signal what we

find important enough to gauge. According to Bearman and Ajjawi (2018), sharing criteria with learners signals that transparency can be achieved, giving the students a "*sense that knowledge is fixed and stable*" (p. 4).

Empirical evidence backing up the claims

The empirical support for the subthemes is mixed. In relation to the steering effect of assessment criteria, most claims are based on anecdotal evidence based on personal experiences or theoretical assumptions. Although references are made to previous publications, these previous publications may in turn be based on personal experiences or theoretical speculations, or they may have a more general focus, not specifically addressing the use of explicit criteria. As an example of the former, Kohn (2006) refers to Wilson (2006), who in turn based her claims mostly on personal experience. As an example of the latter, in the section "Effects of rubrics on learning", Mabry (1999) refers to a study that discusses the effects of state-mandated testing, which is a much broader issue than the use of explicit criteria. Similarly, the "backwash of assessment" (Biggs, 1999), where students strategically adjust their approaches to learning depending on what is being assessed (e.g., Biggs, 1987; Entwistle, & Ramsden, 1983), has been documented in a number of studies (for an overview, see Struyven, Dochy, & Janssens, 2005), but this is neither a recent phenomenon nor specifically linked to the use of explicit criteria. Rather, some students are likely to strategically adjust their learning strategies to assessment demands, as long as there are high-stake consequences attached to assessment outcomes. This is reflected in a study by Bell, Mladenovic, and Price (2013), where students provided their reflections on the usefulness of different resources, such as marking guides, grade descriptors, and annotated exemplars (i.e., not only explicit criteria). In this study, most students found such resources to be useful, but while some students focused on the mechanics of the

assessment task (45%), others used the resources as tools for reflection and learning (50%). This means that not all students adopt an instrumental approach (in this case, a minority of the sample), again suggesting that this is not an effect of using explicit criteria, but a more general phenomenon.

In relation to the extensive use of coaching, Torrance (2007) reports on a series of parallel case studies of assessment in the post-compulsory "Learning and Skills Sector" (e.g. further education colleges, workplaces, and adult learning environments), as well as results from a questionnaire. Findings suggest, for example, that achievement is often defined in narrow and instrumental terms and that there is an overwhelming support for learners at every level, for instance by breaking down and interpreting assessment criteria. Similar to the examples above, although the use of explicit criteria seems to be detrimental for students' long-term learning and autonomy in these situations, it is difficult to extract the impact of using explicit criteria from the wider pedagogical context.

Finally, in relation to the transfer of "epistemic beliefs", although a plausible claim, it still needs to be shown whether the "epistemic beliefs" transferred when sharing explicit criteria differ as compared to the beliefs transferred in other pedagogical situations, but also to what extent such beliefs have an influence on student learning.

Empirical evidence against the claims

In the literature on rubrics, there are several examples of studies, where students use rubrics to: (a) understand expectations, (b) self-regulate their learning, and (c) improve their task performance, which seem to contradict the idea that sharing assessment criteria with learners leads to instrumentalism and surface learning strategies.

a) Andrade and Du (2005) report that students find rubrics useful to understand the teacher's expectations. Reynolds-Keefer (2010) and Jonsson (2014) also report similar findings. Panadero, Alonso-Tapia, and Huertas (2014), who measured students' perceptions of the usefulness of rubrics and scripts (i.e. an instrument that can also be used for self-assessment), show that rubrics clearly outperformed scripts in terms of perceived usefulness. Furthermore, in this study it was found that students using rubrics claimed to be more *focused on learning*, as compared to the students using scripts, and no significant differences were found between the groups for *focusing on grades*. Taken together, this research suggests that rubrics may help students to understand the expectations, without necessarily having negative effects on students' motivational approach to the task, even in some cases promoting learning over performance goals. It is also important to remember that there is almost always a need for students to comply with the teacher's expectations, but with access to rubrics some students seem to feel more confident, experiencing less anxiety, which means that they may become more motivated to focus on learning, without the fear of failure.

b) Regarding self-regulated learning, in Reynolds-Keefer (2010) rubrics were an aid to students in both planning and performing the assignment. Most students claimed to read the course outline and then beginning the assignment, using the rubric as a reference point throughout the process. Similar findings are reported by Jonsson (2014). Another example is a study by Panadero and Romero (2014), where students using a rubric reported higher levels of SRL strategies throughout the three self-regulatory phases (forethought, performance, and self-reflection) as compared to students in other conditions. These studies show that rubrics can support students to become more self-regulating, making them

proactive learners, rather than passive recipients or mindless followers of criteria. Furthermore, the meta-analysis on the effects of self-assessment interventions on self-regulation strategies, by Panadero and colleagues (2017), did not report any significant effects of rubrics on students' self-regulated learning, unless combined with self-assessment interventions. This suggests that rubrics may have better effect on student self-regulation if implemented for formative purposes.

c) Studies presenting evidence of improved student performance can be divided into two different categories: (i) studies where the students are provided with rubrics before task performance, and therefore can use the rubric to plan and monitor their performance, and (ii) studies where the students use rubrics to revise/improve previous task performance (Jonsson, 2020). As an example of (i), Kocakülah (2010) compared the performance of two groups of pre-service primary science teachers; one treatment group, where students were instructed about the features and use of a rubric and also asked to construct a rubric, and one control group without a rubric. Except for the rubric, both groups were taught the same content (Newton's Laws of Motion), using the same teaching methods. Findings show that the students in the treatment group performed significantly better when solving problems after the intervention, as compared to the students in the control group. As an example of (ii), biology students were recorded while giving two oral presentations. After the first presentation, one group of students was required to complete a self-assessment rubric of their presentations, while another group was only encouraged to watch the recordings of themselves. All students also received both instructor and peer feedback. For their second presentations, the students who completed the self-assessment

> rubric received higher scores than the control group on several criteria (Ritchie, 2016). Studies such as these, where student performance is improved when students are provided with a rubric, suggest that the students are able to use the information provided in rubrics in order to improve the quality of their task performance. On the one hand, this could be interpreted as an indication of "criteria compliance", since students obviously improve how well they meet the criteria. What is generally missing, however, is information regarding whether the students were able to meet these criteria in a superficial way (and, as discussed above, whether the rubric is a valid instrument for assessing the performance at hand). If the students are able to improve their performance, without acquiring or using the intended knowledge or skills, it would be a clear case of "criteria compliance". However, if the rubric supports the students in acquiring or using the intended knowledge or skills (i.e. they learn what they were expected to learn), should it still be considered "criteria compliance" in the sense that all intended learning is seen as superficial and conformative, while only the unexpected and unique is seen as an authentic and desirable outcome of education?

Taken together, there is empirical support for the claim that the use of explicit criteria (in the shape of rubrics) may support student learning by helping students to understand expectations, self-regulate, and improve their task performance. In these studies, the students generally act as mindful users of criteria, by balancing the demands for short-term performance on summative assessments on the one hand and long-term learning on the other. As noted above, however, some of the controversy in relation to this question relates not only to the use of rubrics, but to whether we value this compliance with expectations as a desirable outcome of education or not, or whether intended learning

should be seen as superficial and conformative per se. This is fundamentally a question about what we perceive as "good education" (Biesta, 2008), and where we tend to balance the three functions of education (i.e. qualification, socialization, and subjectification) differently.

Theme #3: Simple implementations don't work

This theme explores the lack of effects when rubrics are implemented in a superficial manner, such as just handing out the rubric to the students. The absence of a positive impact may be an effect of students' limited experience of using rubrics (Andrade, 2005), but also teachers' (Lovorn & Rezaei, 2011).

Summary of main criticisms

Three subthemes were identified:

a) Simple interventions do not work (excerpts 18, 19, 81, 82, 83, and 84). These authors argue that just handing rubrics is not enough, either based on their empirical evidence or citing others' work.
b) Teachers need training too. Excerpt 61 shows that untrained teachers may implement the use of rubrics in non-appropriate ways (e.g. relying on them too much), which can be amended through training.

c) No substitution for good instruction and assessment. Excerpts 20, 53, 54, 58, and 59 point out that rubrics cannot, in themselves, replace teaching or make assessment decisions for the teachers.

Empirical evidence backing up the claims

Regarding the first subtheme, only one of the excerpts provides direct empirical evidence to support the claim (O'Donovan et al., 2004, excerpt 18), while another study apparently points in the same direction, but the lack of control group invalidates such a conclusion (Wollenschläger et al., 2016). The rest of the excerpts are either theoretical

and written by a proponent of rubrics (19) or referring to previous work that does not support the claim (81, 82, 83, and 84). It should be noted, however, that excerpt 84 includes two references in German. Regardless of these excerpts, there is literature supporting the claim that simple interventions (e.g. just handing out rubrics) do not always work (e.g. Arter & McTighe, 2000; Brookhart, 2013; Panadero & Jonsson, 2013). However, none of these sources were used in the excerpts.

Regarding the second subtheme, excerpt 61 presents four references. Two of them are anecdotal, another one is a publication in favour of using rubrics, and the last one has validity threats in their interpretation of the results (Rezaei & Lovorn, 2010). Finally, regarding the third subtheme, there is a criticism in excerpt 20 coming from personal experience: “Similarly, rubrics are not a replacement for good instruction. Even a fabulous rubric does not change the fact that students need models, feedback, and opportunities to ask questions, think, revise, and so on” (Andrade, 2005, p. 29). The rest of the excerpts (53, 54, 58, and 59), coming from two articles by Rezaei and Lovorn (2010) and Lovorn and Rezaei (2011), are cases of particular concern as they misrepresent previous research. As shown in Appendix 2, these authors make reference to articles that do not support their claims.

Empirical evidence against the claims

It is commonplace for advocates of the use of rubrics to emphasize the factors pointed out in these subthemes, namely that both students and teachers need training in using rubrics, and that rubrics, no matter how helpful as instructional resources, cannot substitute for high-quality teaching and assessment (Andrade & Valtcheva, 2009; Brookhart, 2018; Panadero & Jonsson, 2013). Therefore, we also believe that some of these claims are logical and should be contemplated when implementing rubrics. However, in this review we are contrasting what is known from empirical evidence and

the vast majority of the excerpts analysed above present problems in this regard. For example, although current empirical research supports the claim that students need training in order to use rubrics productively, it should be acknowledged that there are a number of studies in higher education contexts, where students have managed to use rubrics to improve their performance and/or self-regulate without any substantial training (Cheng & Chan, 2019; Panadero & Romero, 2014). For school students, however, longer and more comprehensive implementations are typically needed in order for the use of rubrics to be effective (Panadero & Jonsson, 2013). Consequently, recommendations for the implementation of the use of rubrics for formative purposes are often accompanied by the implementation of self- and peer assessment (e.g. Andrade, 2005). Interventions have also been made in order to help students internalize assessment criteria, so that their use and understanding of the rubric may become deeper and more advanced (Fraile et al., 2017; Joseph, Rickett, Northcote, & Christian, 2019).

Theme #4: Limitations of criteria and analytical assessments

Criteria are a fundamental building block of rubrics, which means that if criteria have severe limitations, then the use of rubrics will be affected by these limitations. In this category, we have gathered criticisms based on the proposed limitations of criteria and analytic assessments.

Summary of main criticisms

Four main subthemes can be identified:

a) There is no precision in criteria. It is a common assertion that rubrics (or criteria) can never deliver precision in assessment (excerpts 26, 52, 62, 63, 64, 79, 80). Using criteria therefore only provide an illusion of accuracy, while the assessment is in fact left to the individual teacher’s discretion (Kohn, 2006).

b) Analytic assessments of individual criteria are not valid (excerpts 7, 14, 39, 40, 41, 42, 47, 56, 74, 75, 85, 86). There are two dimensions of this subtheme, where the first recognizes that analytic assessments do not reflect the actual assessment process, since according to Sadler (2009), assessors are "more interested in how the work comes together as a whole than in performance on individual criteria" (p. 167). Second, Sadler (2009) claims that when analytic and global judgments differ, teachers may not always be able to account for the discrepancy because the "work exhibits an indefinable 'quality,' inherent in its wholeness, which simply cannot be passed over as irrelevant or inconsequential." (p. 166).

c) No list of criteria is complete (excerpts 42, 43, 44, 45, 88). There are two dimensions of this subtheme as well, and the first recognizes that in most contexts, it is not possible to use all conceivable criteria, which means that some criteria have to be selected, while others are excluded. At the same time, a certain criterion, which was not included in the preset list distributed to students, may turn out to be crucial to the judgment (Sadler, 2009, p. 166). Second, Sadler (2009) claims that it is a necessary condition that the criteria are conceptually distinct from one another: "Each criterion is assumed to have an established interpretation that, at least in theory, represents a property that is different from those signified by the other criteria, taken singly or together." (p. 166-167).

d) Some criteria are not possible to articulate. According to Sadler (2009), some evaluative characteristics can only be "tacitly known" and therefore not only difficult to articulate, but actually impossible (excerpts 51, 73).

Empirical evidence backing up the claims

Most of the criticism on criteria is presented by Sadler (2009; 2014) and several other critics cite his work. His arguments are, however, mostly based on anecdotal evidence based on personal observations, which have "…been investigated in conversations with academics, then elaborated and refined." (Sadler, 2009, p. 164), while few are empirically investigated. Others refer to publications such as Kohn (2006), Chapman and Inman (2009), or Moskal and Leydens (2000), none of which are scientific empirical studies.

One of the few empirical studies presenting findings that support the claims in this category is a study by Bloxham, Boyd, and Orr (2011), where twelve lecturers from two universities were asked to think aloud as they graded two written assignments. It was found that the assessors made holistic rather than analytical judgements and that a number of assessors did not make use of written criteria at all. When criteria were used, it was a post-hoc process, justifying a holistic decision already made. This study therefore supports the idea that analytic assessments do not reflect the actual assessment process of teachers, as claimed by Sadler (2009).

Empirical evidence against the claims

In the literature on rubrics, there are studies investigating: (a) the reliability of assessments supported by rubrics, including (b) studies that compare analytic and holistic assessments. There are also examples of research, where (c) tacit criteria have been identified and articulated.

a) Two reviews on research about rubrics have reported on the reliability of assessments made with the aid of rubrics. First, Jonsson and Svingby (2007), who reviewed 75 studies, concluded that rubrics seem to aid assessors in achieving high internal consistency when scoring performance tasks. This is logical, since the same assessor is likely to make similar interpretations of the

criteria at different occasions. However, they also concluded that the majority of studies reporting on assessor consensus did not exceed 70 percent agreement, which is often considered the lowest acceptable level. In a later review, Brookhart and Chen (2015), who reviewed 63 studies, concluded that assessors using rubrics can achieve acceptable levels of consistent and reliable judgment, even if there are exceptions. Taken together, current research suggests that rubrics may support acceptable levels of reliable assessment, particularly intra-rater agreement, although improvements can be made through the clarity of the rubrics and training of the assessors. In particular, there is a contrast between rubric-supported assessments, and situations where no shared criteria are used. The latter case is common in early studies on teacher assessments, where reliability coefficients often were little better than sheer guesses (Brookhart et al., 2016; Parkes, 2013). This suggests that criteria may not have the same precision as numerical measures, but that precision is greater when using rubrics, as compared to judgments without the aid of such instruments.

b) Although analytic assessments may not be the natural choice for teachers, this does not necessarily disqualify such a procedure. In the review by Jonsson and Svingby (2007), analytical assessments are shown to be preferable for enhancing the consistency of the assessment. Similar results are presented in an experimental study by Jonsson and Balan (2018), where there is a substantive difference in terms of agreement between teachers using analytic (66%) or holistic (46%) grading. Most importantly, however, there were no differences between the groups in terms of justifications for the grades, which suggests that analytic and holistic grading did not differ in terms of validity. In situations where there is a need for increased reliability, such as when grading, it would

therefore seem that analytic assessment is preferable, even if not the traditional mode of assessment for teachers. Furthermore, as shown by Tomas, Whitt, Lavelle-Hill, and Severn (2019), holistic assessments do not necessarily align with what assessors assume they are assessing.

c) Prominent in the field of investigating the tacit knowledge of assessors is the work by Lars Lindström. In one of his studies, criteria were articulated for handicraft work, by using a method known as "repertory grids" (Lindström, 2008). This method is well suited to extract and articulate parts of the tacit knowledge within a community of practice, and in this particular study, five categories of assessment criteria were formulated, which were shown to be used by experts in the field of handicraft. In another study, Lindström (2006) created a scoring rubric for assessing creativity. As opposed to the inductive design in the study on handicraft criteria, the criteria in the rubric for creativity were formulated on the basis of previous research and current discourse (for example as expressed in journals), and then tested by art teachers. The work by Lindström therefore suggests that it is indeed possible to articulate the "tacit knowledge" of assessors.

Taken together, there is empirical support for the claim that the use of rubrics increases the reliability of assessments, and that analytic assessments are sometimes superior to holistic assessments in terms of consistency, without necessarily sacrificing validity. This means that, when there is a discrepancy between analytic and holistic assessments, the holistic assessment is not necessarily more accurate. As suggested by Tomas et al. (2019), analytic and holistic assessments may partly capture different things and should therefore not be seen as opposites, but rather as complementary. There is also empirical support for the claim that tacit criteria may be identified and articulated.

Theme #5: Context dependence

The question guiding this theme is whether it possible to understand and use the same rubric in different contexts, or whether rubrics are context specific. The criticism is mostly based on a discussion about the context-dependency of assessment criteria in general, as outlined by Sadler (2009), and not specifically about rubrics.

Summary of main criticisms

Two subthemes were identified:

a) Assessment criteria are context dependent. This is the main claim that can be found in excerpts 48, 49, 50, 87, and 93. According to Sadler (2009), criteria can be interpreted "differently by different teachers…or differently by the same teacher in different contexts" (p. 169). Similarly, Bearman and Ajjawi (2018) argue that academic standards and criteria are socially constructed and therefore context dependent.

b) Two excerpts included claims that were directly related to rubrics. Ito (2004, excerpt 76), citing Shay (2004), concludes that rubrics are context-dependent and, Humphry and Heldsinger (2014, excerpt 71) point out that rubrics can be content specific or general, and that it is the instructional decision behind the use of rubrics that determines whether a broader applicability is desirable.

Empirical evidence backing up the claims

These claims are theoretically grounded by the authors, some of them heavily based on anecdotal evidence, with no references to empirical evidence. This is the case even for the publications including empirical evidence for the use of rubrics (e.g. Brookhart, 2018; Humphry & Heldsinger, 2014).

Empirical evidence against the claims

In relation to this theme, there is a tension between the different purposes of assessment. For example, while proponents for large-scale and high-stakes assessments are likely to view assessment criteria as general and independent of context, and thereby possible to use across different schools, regions etc., from a classroom assessment perspective, criteria need to be tailored by the teachers for each particular learning context. A similar distinction is made by Ajjawi and Bearman (2018), who contrast a representational view of criteria (and/or standards), which assumes that a criterion is an accurate and stable representation of something, and that this "something" is separate from the knower, with a sociocultural view, in which the context and its social and cultural relations are taken into account. In relation to the former, representational view, criteria are more or less easily transferred across contexts, since each criterion has one single meaning, which does not change in relation to the context or the person who interprets them. In the latter, sociocultural view, explicit criteria are only "the tip of the iceberg" (Holmstedt et al., 2018), while the greater part is tacit, residing in the practices of academic, and professional communities (O'Donovan et al., 2004). Taken together, it is difficult to claim that rubrics are either context-dependent or context-independent, since it depends on the purpose of the assessment and the perspective taken.

Another tension is whether rubrics should be general or specific, which depends on the instructional goals and the purpose for using rubrics (Brookhart, 2013; 2018). If a rubric is going to be used state-wide, it will have to be created with a general approach, including the main characteristics of the task at hand. However, if a teacher wants the students to train for a particular skill, the accompanying rubric would typically be more specific and less transferable. The question of whether rubrics should be general or specific is therefore an instantiation of the continual need to carefully and responsibly design the assessment tools used.

Theme #6: Miscellanea

This "theme", which is not a coherent or defined theme, includes points of criticism that do not fit into the main themes. Each of these points of criticism are typically voiced by one author only.

Summary of main criticisms

Three points of criticism can be identified:

a) According to Kohn (2006, excerpt 24), the use of grades should be abandoned. Since rubrics legitimate grades, they should also be abandoned.
b) Rubrics are time-consuming to create, which could be stressful for both teachers and students. Rubrics should therefore only be developed for the most important and complex assignments (Ito, 2015, excerpt 77).
c) Per Bearman and Ajjawi (2018, excerpt 89), transparency can be conceptualized as contributing to a system that seeks to commodify and control education. Explicit criteria thereby allow institutions to control teachers and teaching.

Empirical evidence backing up the claims and empirical evidence against the claims

The criticism by Kohn (2006) is based on values and can therefore not be supported or refuted by empirical data. Neither can the criticism by Bearman and Ajjawi (2018). However, both points of criticism are linked to certain assumptions, for instance that rubrics are used for summative purposes (only), and that transparency is a political, rather than a pedagogical, tool.

That rubrics may be time-consuming to create can be perceived as more or less self-evident and is supported by empirical studies (e.g. Rochford & Borchert, 2011). However, there is also empirical support for the claim that rubrics – when ready – can save time during the assessment process (e.g. Halonen, Bosack, Clay, & McCarthy,

2003; Helvoort, 2010). Therefore, the time invested in creating a rubric may be compensated in the longer run.

Discussion

The aim of this review was to investigate the criticism voiced against the use of rubrics by identifying the main themes in this criticism, investigating the empirical evidence backing the claims made by critics, and (if available) presenting empirical evidence against the claims.

From the 27 publications included in the review, 93 excerpts were identified and organized around 5 main themes and one "theme" including points of criticism that do not fit into the main themes. The themes have been named: a) Standardization and narrowing the curriculum, b) Instrumentalism and "criteria compliance", c) Simple implementations don't work, d) Limitations of criteria, e) Context dependence, and f) Miscellanea.

The themes have a varying degree of empirical evidence to support the claims made. In some cases, such as the theme *Simple implementations don't work,* the criticism is clearly backed up by empirical research. There are also a smaller number of studies that provide empirical support for some of the claims, such as Bouwer et al. (2018), showing that the use of rubrics may narrow some dimensions of peer feedback, as compared to using comparative judgment for communicating standards. Another example is the study by Humphry and Heldsinger (2014), which showed that the use of equal numbers of performance levels for all criteria may produce a halo effect. However, overall the criticisms are largely either based on anecdotal evidence, theoretical speculations or extrapolated from broader issues, such as large-scale and high-stakes accountability testing.

Regarding the counter arguments, in some cases, such as whether *Rubrics are context dependent*, the tension cannot be resolved by empirical evidence, because it is a matter of purpose and perspective. Similarly, some of the claims in the *Miscellanea* theme are based on values, which can therefore not be countered with arguments based on empirical evidence. In other cases, the tension is difficult to resolve, because there is a lack of empirical evidence, such as studies investigating the validity of rubrics. Importantly, in the opposed direction there are many empirical studies providing evidence that can be used to counter the critical claims in most of the themes. These findings are discussed below.

Theoretical concerns and over-generalizations

It becomes clear from the studies reviewed, that there are a number of concerns and potential dangers with using and sharing assessment instruments with explicit and pre-set criteria, such as rubrics. It is claimed, for instance, that rubrics will narrow the curriculum and lead to instrumental learning. As shown by some studies, these claims may indeed turn out to be true, at least under certain circumstances or for certain students (Bouwer et al., 2018; Torrance, 2007). What is problematic, however, is the determinism and generality of the claims. This is most evident in the work by Sadler (2009; 2014), who claims, for instance, that *any* attempt to "codify" academic standards is futile, that it is a *necessary* condition that criteria are conceptually distinct from one another, or that some criteria can *only* be "tacitly known" and *impossible* to articulate. But even if students do not fully understand the criteria, or the quality levels do not perfectly capture the standards,, rubrics can still facilitate a dialogue between the students and the teacher about what the criteria mean, thus beginning the process of enculturating the student into a community of practice of which the teacher is already a member. Additionally, since the position is dogmatic, it effectively closes the door for

empirical investigations. A more reasonable position would be that it is *difficult* to codify academic standards (and therefore, may not be worthwhile) and/or that these codifications *only partly* capture the complexity of the "true standard" (e.g. Rust, Price, & O'Donovan, 2003). Such a less dogmatic position is more reasonable because it opens for the possibility to investigate, for instance, which standards that can be communicated through codifications and which cannot, which dimensions of these standards that need to be supported by other modes of communication, how different modes of communication can contribute to a more complex understanding of the standards, and so on.

Based on the findings from this review, we would therefore like to suggest a more pragmatic approach, where potential limitations of rubrics are investigated empirically through scientific inquiry and decisions are based on such data. An excellent example is the theoretically grounded claim that it is not possible to achieve transparency in assessment through explicit criteria (e.g. Ajjawi & Bearman, 2018). Still, there are several empirical studies reporting that students actually use criteria to regulate their learning and improve their performance (e.g. Brookhart, 2018; Panadero & Jonsson, 2013), which seem to contradict the claim that students are not able to understand or make use of explicit criteria.

Assuming the worst

The US has a long history of using large-scale tests for high-stakes accountability purposes (Nichols & Harris, 2016), and accompanying rubrics, for assessing students' writing performance (Turley & Gallagher, 2008). In some sense, it is therefore not surprising that rubrics are associated with large-scale assessments and that a number of critics assume that rubrics can only be used for high-stakes and/or other summative assessment situations, such as grading. In line with this assumption, and as

reflected by the themes identified in this review, these critics typically argue that rubrics: (a) focus on surface details in order to increase reliability, (b) include quantifiable levels of performance, and (c) that assessments of individual criteria are summarized into aggregated scores.

While the features mentioned above may very well characterize rubrics used for accountability purposes, this is not necessarily true for the formative use of rubrics. For example, if the primary purpose of using rubrics is to communicate expectations to students, reliability is not the major concern. Rubrics used for such purposes can therefore include dimensions that are considered important, even if this means using more “fuzzy” criteria, rather than easy-to-agree-upon criteria, because the rubric is not used in “one-shot” high-stakes situations. Furthermore, the rubric does not stand alone, but can be supplemented with other modes of communication.

Similar to the alleged focus on surface details, rubrics used for formative purposes do not need to include quantifiable levels of performance. On the contrary, according to Brookhart (2018), rubrics used for formative purposes should focus on qualities and not quantities, so that they are able to provide guidance to students (and to teachers providing feedback to students). This also means that the outcome of the assessment should not be summarized into an aggregated score but communicated as strengths and weaknesses in relation to different criteria – which is the raw material for formative feedback.

Taken together, a distinction has to be made between formative and summative purposes of rubrics, as well as between the assessment of student performance and school evaluation purposes, not only in relation to how they are used, but also in relation to how they are designed. As noted in the introduction, rubric design is affected by the purpose of assessment and while high reliability may require task-specific rubrics

with fewer levels, other designs may be more adequate for formative purposes. However, the limited empirical evidence about rubric design has still to test this type of hypothesis.

Are rubrics representational or socio-cultural?

An important question for a number of critics is whether it is possible to move rubrics across different contexts. As briefly mentioned above, this is a complex question, since the answer depends on which perspective is taken. One dimension of this question is whether we adopt what Ajjawi and Bearman (2018) refer to as a *representational* or a *socio-cultural* view of criteria. In the former case, it is assumed that the codifications are more or less fixed representations of the qualities they signify, which makes possible transfer across time, geography, and disciplines (cf. Stake, 2004). Although this might be true for some extraordinarily stable pairs of codification-quality, in most cases the relationship is less static. This is sometimes easier to see in relation to practices that are unfamiliar or obscure to us, as the art of wine tasting or martial arts may be to some. In both of these practices, the quality of "balance" is essential. However, it is quite obvious that the codification (i.e. "balance") does not refer to the same trait. It is also obvious that in order to understand the quality, you have to have an understanding of the practice it belongs to, because the "real quality" is tacit; hidden in the practice of the practitioners. This does not, however, mean that it is necessarily *impossible* to articulate and communicate these qualities verbally.

A socio-cultural view of criteria makes clear that rubrics are context-dependent in the sense that they are linked to the language and practice of a specified community of practice. Such a view also implies that the relationship between codifications and qualities is not fixed, but dynamic, as it may change due to changes in language or practice. If rubrics are used within this community of practice, there is no need for

"universal rubrics" or transformations across contexts, where the meaning of criteria is "lost in translation". Instead, the rubric becomes one of several tools for learning to identify and appreciate the qualities of the community. This is clearly exemplified by Holmstedt et al. (2018), where preservice teachers learned how to discern aspects of teaching that had previously been concealed to them, because they did not know what to look for – and this discernment was facilitated by the access to explicit criteria.

Using a map or finding your own way

Biggs (1999) once wrote that:

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

Assessment has an impact on some students' approaches to learning, regardless of whether we use rubrics or not, as this has been shown in studies well before the era of rubrics (e.g. Säljö, 1975). The issue of strategic learning is therefore not confined to the use of explicit criteria, although it cannot be ruled out that sharing criteria could possibly reinforce such learning. Whether the steering effect of assessment is something good or bad is, however, a matter of debate. Of course, if taking the view that rubrics only include superficial and easy-to-measure performances, then it is surely not a good thing. But assuming that the rubric is valid and "mirrors the curriculum", and students learn what they ought to be learning, is it still a bad thing? Some, like Biggs (1999), would say "no", others would still say "yes". Torrance (2012), for instance, writes:

> Are we trying to get students to jump through pre-specified hoops, by making the nature of those hoops more apparent and encouraging students

> to better understand how the objectives of a course can be met; or are we trying to get students to think for themselves? (p. 330)

What Torrance (2012) opposes is students receiving too much support, because if guided too heavily, the students become passive and the challenge of learning is removed. The idea of instrumentalism is therefore not only about whether rubrics are valid and mirror the curriculum, but also about being too supportive when sharing criteria with learners. This is in some sense a true dilemma. On the one hand, highly detailed rubrics may not leave enough space for creative and divergent thinking, but, on the other hand, relying on the transfer or tacit knowledge, guild knowledge, or connoisseurship means that we run the risk of making students more dependent on their teacher rather than the other way around. If the assessment criteria are concealed from the students, they are deprived of the possibility to take full responsibility for their work (Balloo, Evans, Hughes, Zhu, & Winstone, 2018). Therefore, a middle way is needed, where students are aware of the conditional nature and indeterminacy of criteria (Sadler, 2009), and where criteria are indicators of quality, without dictating exactly what students should do, or how they should do it (Wiggins, 1998).

This middle way can be approached in different ways. One would be a scaffolded approach: providing support while it is needed and then gradually take it away (e.g. Taylor, 1911). This type of approach was proposed by Panadero (2011), using two different tools (rubrics and scripts), hypothesizing that their combined effect would counteract an excessive dependence on the tools and make further contributions to students' understanding of the criteria. Another approach is the co-creation of rubrics, which involves the students in the selection of criteria and development of rubrics (Fraile, Panadero & Pardo, 2017). This approach seems to gain some momentum, with at least three publications during the last year (Bacchus et al., 2019; Joseph et al., 2019;

Kilgour et al., 2019). Finally, Jonsson and Eriksson (2019) used criteria for the assessment of students' argumentation in astronomy, where the arguments should be based on data and there should be more than one observation in favour of the claim/conclusion. For higher quality, counterarguments should be presented and answered, and arguments should be backed with scientific knowledge. Furthermore, for high quality, qualifiers could be addressed. Based on these criteria, the students peer-assessed each other's arguments and provided feedback. These criteria did not tell the students exactly what to do, or how they should do it. Rather, the criteria gave the students the possibility to engage more deeply with the academic content, as they did not have to focus on the "rules of the game".

Taken together, whether the use of rubrics leads to instrumental learning is not a simple black-or-white question, but a question with several shades of grey. First, not all students are affected similarly (e.g. Bell et al., 2013). Second, strategic learning is not confined to situations where rubrics are used but has been observed in a number of different contexts well before the use of rubrics became commonplace (e.g. Biggs, 1987). Third, rubrics can to a greater or lesser extent be valid and mirror the curriculum depending in their design and implementation (Brookhart, 2018). Fourth, the criteria in the rubrics can to a lesser or greater extent specify exactly what the students should do, and how they should do it (Wiggins, 1998). Fifth, if students are not guided by explicit criteria, they have to be guided by something else, such as trial-and-error attempts or cues from the teachers, which can make it difficult for the students to take responsibility for and regulate their learning (Balloo et al., 2018; Panadero & Jonsson, 2013). This is particularly true for the position taken by Wilson (2006), who suggests not only that assessments should be based on what Sadler (2009) calls "emergent criteria" (i.e. criteria that surface while evaluating student performance), but also on the individual

reactions evoked by student performance. Taking such a position means, first, that students are not given access to the assessment process, and cannot take control of their learning, since the criteria are not disclosed. Instead, they have to rely on their teacher's judgment, until they have been socialized into the practice safeguarded by the teacher. Second, it means that students may be assessed in relation to different criteria, since which criteria are "activated" depends on the reactions of the teachers. This, however, could be argued to run against fundamental principles of formative assessment, which suggest that assessment criteria and learning intentions should be made available to the students, so that they can become owners of their own learning processes (e.g. Black & Wiliam, 1998).

Limitation and future work

This review has an important limitation to bear in mind. While we performed a systematic literature search, including three methods for finding publications critiquing rubrics, we did not perform a similar search for literature in relation to the "empirical based responses" to these criticisms. Rather, we used our own knowledge and reference libraries to write the counterarguments.

Regarding future work to be done in this area, we want to issue a call for forthcoming articles about rubrics about some "predatory" use of references. As we have exposed in our data, there is a large number of publications making claims against rubrics based on other scholars' references that do not support the claims. Cases range from citing anecdotal evidence as if it was scientific empirical evidence, to using rather positive ("pro-rubrics") publications as if they were negative towards these tools. It seems like some of these cases are disingenuous uses of other scholars' references. Importantly, we suggest that when a colleague wants to critique rubrics based only on anecdotal evidence based on personal experiences, not using any type of scientific

inquiry, this should be explicitly stated upfront. Additionally, we have already suggested a more pragmatic approach, where potential limitations of rubrics are investigated empirically and decisions are based on such data, instead of relying on anecdotal data and/or theoretical speculation. This is the direction taken by, for instance, Humphry and Heldsinger (2014) and Bouwer et al. (2018), which are some of the few papers presenting empirical research to explore the limitations of rubrics. We suggest that this is the way to follow, since rubrics (similar to all pedagogical tools) have limitations and most of these limitations have not been sufficiently explored. We also suggest that theoretical criticism would be more helpful if not assuming that rubrics per definition focus on surface details, are used for accountability purposes, or are otherwise inadequately designed or implemented. Instead, a more nuanced and less deterministic conceptualization of rubrics is needed in order to advance our understanding of both benefits and limitations of rubrics.

Finally, an important direction for future research is to address the distinction between performance and learning (see Soderstrom & Bjork, 2015). While rubrics have been shown to support the improvement of student performance, it still remains to be explored whether they also support student learning in a wider sense, for instance in terms of long-term retention or transfer to other tasks. In the light of this review, this distinction seems to be crucial for the advancement of the research about the educational consequences of rubrics.

Conclusions

One of the main conclusions from this review is that the empirical evidence supporting the claims of the critics is, with only a few exceptions, neither direct nor strong. First, some critics refer to anecdotal evidence and/or own personal experiences, which have limited value as scientific evidence. More importantly, only a few of these

critics clearly state the nature of the data, while others even write as if their claims were backed up by strong scientific and empirical evidence. Second, there are a number of misreads or/and misrepresentations of previous research.

Our review shows that the relationship between theoretically grounded concerns and empirical findings has not been "nurtured" in the criticism of rubrics. Some of the critics present a significant number of concerns about rubrics and anticipate negative consequences based on theoretical considerations. Even if some of these considerations are legitimate, the criticism is sometimes based on an overly deterministic position. We therefore propose a pragmatic approach, where we investigate what actually happens when using rubrics, and base our decisions on empirical data.

Furthermore, a number of critics make assumptions about rubrics with a narrow conceptualization of rubrics in mind. One prevalent assumption is that rubrics are only used for accountability testing and/or other summative assessment situations, such as grading, when we know that is not the case (e.g. Panadero & Jonsson, 2013). Other common assumptions are that rubrics are designed to be universal and transferable across contexts (i.e. based on a representational view of criteria), that rubrics specify exactly what to do and how, that rubrics (or analytic assessments) are not valid or mirror the curriculum, and that rubrics cannot be supplemented with other means for communicating expectations. Since it is clearly not fair to judge the merits of a rubric designed for formative use in the classroom, based on the assumption that rubrics can only be used for large-scale assessments, this narrow conceptualization of rubrics needs to be replaced with a broader understanding, which may encompass rubrics designed and used for different purposes, both formative and summative.

In sum, as with any instructional or assessment tool, rubrics can be used differently and for different purposes. Similar to oral exams, multiple-choice tests, and

other assessment tools, rubrics may be ill-designed and/or poorly implemented. Even if well designed, a poor implementation is likely to reduce the benefits of using a rubric; as is a poorly designed rubric, even if well implemented. However, we will only optimize the design and implementation of rubrics through scientific empirical research on benefits, as well as limitations.

References

*The publications indicated with * were included as empirical evidence in this review.*

Ajjawi, R., & Bearman, M. (2018). Problematizing "standards": representation or performance? In D. Boud, R. Ajjawi, P. Dawson, & J. Tai (Eds.), *Developing evaluative judgement in higher education: Assessment for knowing and producing quality work* (pp. 41-50). Oxon; New York, NY: Routledge.

*Andrade, H. (2005). Teaching with rubrics: The good, the bad, and the ugly. *College Teaching, 53*(1), 27-32. doi:10.3200/CTCH.53.1.27-31

*Andrade, H. L. (2006). The trouble with a narrow view of rubrics. *English Journal*, 9-9.

Andrade, H., & Du, Y. (2005). Student perspectives on rubric-referenced assessment. *Practical Assessment, Research & Evaluation, 10*(3), 1-11. Retrieved from http://pareonline.net/getvn.asp?v=10&n=3

Andrade, H., & Valtcheva, A. (2009). Promoting learning and achievement through self-assessment. *Theory Into Practice, 48*(1), 12-19. doi:10.1080/00405840802577544

Arter, J., & McTighe, J. (2000). *Scoring rubrics in the classroom: Using performance criteria for assessing and improving student performance*: Corwin Press.

Au, W. (2007). High-stakes testing and curricular control: a qualitative metasynthesis. *Educational Researcher, 36*, 258-267. doi: 10.3102/0013189X07306523

Bacchus, R., Colvin, E., Knight, E. B., & Ritter, L. (2019). When rubrics aren't enough: Exploring exemplars and student rubric co-construction. *Journal of Curriculum and Pedagogy*, 1-14. doi:10.1080/15505170.2019.1627617

Balloo, K., Evans, C., Hughes, A., Zhu, X., & Winstone, N. E. (2018). Transparency isn't spoon-feeding: How a transformative approach to the use of explicit assessment criteria can support student self-regulation. *Frontiers in Education, 3*(69). doi:10.3389/feduc.2018.00069

*Bearman, M., & Ajjawi, R. (2018). From "Seeing Through" to "Seeing With": Assessment Criteria and the Myths of Transparency. *Frontiers in Education, 3*(96). doi:10.3389/feduc.2018.00096

*Bell, A., Mladenovic, R., & Price, M. (2013). Students' perceptions of the usefulness of marking guides, grade descriptors and annotated exemplars. *Assessment & Evaluation In Higher Education, 38*(7), 769-788. doi:10.1080/02602938.2012.714738

Biesta, G. (2008). Good education in an age of measurement: on the need to reconnect with the question of purpose in education. *Educational Assessment, Evaluation and Accountability, 21*, 33-46.

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

Biggs, J. (1999). What the student does: Teaching for enhanced learning. *Higher Education Research & Development, 18*(1), 57-75. doi:10.1080/0729436990180105

Black, P., & Wiliam, D. (1998). Assessment and classroom learning. *Assessment in Education: Principles, Policy and Practice, 5*(1), 7-73. doi:10.1080/0969595980050102

*Bloxham, S., Boyd, P., & Orr, S. (2011). Mark my words: The role of assessment criteria in UK higher education grading practices. *Studies in Higher Education, 36*(6), 655-670. doi:10.1080/03075071003777716

*Bouwer, R., Lesterhuis, M., Bonne, P., & De Maeyer, S. (2018). Applying criteria to examples or learning by comparison: Effects on students' evaluative judgment and performance in writing. *Frontiers in Education, 3*(86). doi:10.3389/feduc.2018.00086

Brookhart, S. M. (2013). *How to create and use rubrics for formative assessment and grading*. Virginia, USA: ASCD.

*Brookhart, S. M. (2018). Appropriate criteria: Key to effective rubrics. *Frontiers in Education, 3*(22). doi:10.3389/feduc.2018.00022

Brookhart, S. M., & Chen, F. (2015). The quality and effectiveness of descriptive rubrics. *Educational Review, 67*(3), 343-368. doi:10.1080/00131911.2014.929565

Brookhart, S. M., Guskey, T. R., Bowers, A. J., McMillan, J. H., Smith, J. K., Smith, L. F., . . . Welsh, M. E. (2016). A century of grading research. *Review of Educational Research, 86*(4), 803-848. doi:10.3102/0034654316672069

Brown, G. T. L. (2018). *Assessment of student achievement*. New York: Routledge.

*Chapman, V. G., & Inman, M. D. (2009). A conundrum: Rubrics or creativity/metacognitive development? *Educational Horizons*, 198-202.

Cheng, M. W. T., & Chan, C. K. Y. (2019). An experimental test: Using rubrics for reflective writing to develop reflection. *Studies In Educational Evaluation, 61*, 176-182. doi:https://doi.org/10.1016/j.stueduc.2019.04.001

Dawson, P. (2017). Assessment rubrics: Towards clearer and more replicable design, research and practice. *Assessment & Evaluation In Higher Education*, 1-14. doi:10.1080/02602938.2015.1111294

Entwistle, N. J., & Ramsden, P. (1983). *Understanding student learning*. London & Canberra: Croom Helm.

Fraile, J., Panadero, E., & Pardo, R. (2017). Co-creating rubrics: useful or waste of time? The effects of establishing assessment criteria with students on self-regulation, self-efficacy and performance. *Studies In Educational Evaluation, 53*, 69-76. doi:10.1016/j.stueduc.2017.03.003

Greenberg, K. P. (2015). Rubric use in formative assessment: A detailed behavioral rubric helps students improve their scientific writing skills. *Teaching of Psychology, 42*(3), 211-217. doi: 10.1177/0098628315587618

Halonen, J. S., Bosack, T., Clay, S., McCarthy, M., Dunn, D. S., Hill Iv, G. W., . . . Whitlock, K. (2003). A rubric for learning, teaching, and assessing scientific inquiry in psychology. *Teaching of Psychology, 30*(3), 196-208. doi:10.1207/s15328023top3003_01

Helvoort, J. van (2010). A scoring rubric for performance assessment of information literacy in Dutch Higher Education. *Journal of Information Literacy, 4*(1), 22-39. doi: http://dx.doi.org/10.11645/4.1.1256

Holmstedt, P., Jonsson, A. & Aspelin, J. (2018). Learning to see new things: using criteria to support pre-service teachers' discernment in the context of teachers' relational work. *Frontiers in Education: Assessment, Testing and Applied Measurement, 3* (54). doi: https://doi.org/10.3389/feduc.2018.00054.

*Hudson, J., Bloxham, S., den Outer, B., & Price, M. (2017). Conceptual acrobatics: talking about assessment standards in the transparency era. *Studies in Higher Education, 42*(7), 1309-1323. doi:10.1080/03075079.2015.1092130

*Humphry, S. M., & Heldsinger, S. A. (2014). Common structural design features of rubrics may represent a threat to validity. *Educational Researcher, 43*(5), 253-263. doi:10.3102/0013189x14542154

*Ito, H. (2015). Is a rubric worth the time and effort? Conditions for its success. *International Journal of Learning, Teaching and Educational Research, 10*(2).

*Jones, L., Allen, B., Dunn, P., & Brooker, L. (2016). Demystifying the rubric: A five-step pedagogy to improve student understanding and utilisation of marking criteria. *Higher Education Research & Development*, 1-14. doi:10.1080/07294360.2016.1177000

Jonsson, A. (2014). Rubrics as a way of providing transparency in assessment. *Assessment & Evaluation In Higher Education*, 1-13. doi:10.1080/02602938.2013.875117

Jonsson, A. (2020). Rubrics as a tool for self-regulated learning. In P. Grainger & K. Weir (Eds.), *Assessment Rubrics in Higher Education* (pp. 25-40). Newcastle upon Tyne: Cambridge Scholars Publishing.

Jonsson, A., & Balan, A. (2018). Analytic or holistic: A study of agreement between different grading models. *Practical Assessment Research & Evaluation, 23*(12). Retrieved from http://pareonline.net/getvn.asp?v=23&n=12

Jonsson, A., & Eriksson, U. (2019). Formative assessment in Higher Education: An example from Astronomy. In H. L. Andrade, R. E. Bennett, & G. J. Cizek (Eds.), *Handbook of formative assessment in the disciplines* (pp. 146-169). New York, NY, & London: Routledge.

Jönsson, A. & Panadero, E. (2016). The use and design of rubrics to support AfL. In D. Carless, S. Bridges, C. Chan & R. Glofcheski (Eds.), *Scaling up assessment for learning in higher education* (pp. 99-111). Dordrecht: Springer.

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

Joseph, S., Rickett, C., Northcote, M., & Christian, B. J. (2019). 'Who are you to judge my writing?': Student collaboration in the co-construction of assessment rubrics. *New Writing*, 1-19. doi:10.1080/14790726.2019.1566368

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

Kilgour, P., Northcote, M., Williams, A., & Kilgour, A. (2019). A plan for the co-construction and collaborative use of rubrics for student learning. *Assessment & Evaluation In Higher Education*, 1-14. doi:10.1080/02602938.2019.1614523

Kocakulah, M. S. (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

*Kohn, A. (2006). The trouble with rubrics. *English Journal, 95*(4), 12-15.

Lindström, L. (2006). Creativity: What is it? Can you assess it? Can it be taught? *Journal of Art & Design Education, 25*(1), 53-66. doi: 10.1111/j.1476-8070.2006.00468.x

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

Lipnevich, A. A., McCallen, L. N., Miles, K. P., & Smith, J. K. (2014). Mind the gap! Students' use of exemplars and detailed rubrics as formative assessment. *Instructional Science, 42*(4), 539-559. doi:10.1007/s11251-013-9299-9

*Lovorn, M. G., & Rezaei, A. R. (2011). Assessing the assessment: Rubrics training for pre-service and new in-service teacher. *Practical Assessment, Research & Evaluation, 16*(16).

*Mabry, L. (1999). Writing to the rubric: Lingering effects of traditional standardized testing on direct writing assessment. *Phi Delta Kappan, 80*(9), 673-679.

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

Minarechová, M. (2012). Negative impacts of high-stakes testing. *Journal of Pedagogy, 3*(1), 82-100. doi: 10.2478/v10159-012-0004-x

Moskal, B. M., & Leydens, J. A. (2000). Scoring rubric development: Validity and reliability. *Practical Assessment, Research & Evaluation, 7*(10). Retrieved from http://PAREonline.net/getvn.asp?v=7&n=10

Nichols, S. L., & Harris, L. R. (2016). Accountability assessment's effects on teachers and schools. In G. T. L. Brown & L. R. Harris (Eds.), *Handbook of Human and Social Conditions in Assessment* (pp. 40-56). New York: Routledge.

*Norton, L. (2004). Using assessment criteria as learning criteria: A case study in psychology. *Assessment & Evaluation In Higher Education, 29*(6), 687-702. doi:10.1080/0260293042000227236

*O'Donovan, B., Price, M., & Rust, C. (2004). Know what I mean? Enhancing student understanding of assessment standards and criteria. *Teaching in Higher Education, 9*(3), 325-335. doi:10.1080/1356251042000216642

Panadero, E., Alonso-Tapia, J., & Huertas, J. A. (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

Panadero, E., Alonso-Tapia, J., & Huertas, J. A. (2014). Rubrics vs. self-assessment scripts: Effects on first year university students' self-regulation and

performance. *Infancia y Aprendizaje: Journal for the Study of Education and Development, 37*(1), 149-183. doi:10.1080/02103702.2014.881655

Panadero, E., & Jonsson, A. (2013). The use of scoring rubrics for formative assessment purposes revisited: A review. *Educational Research Review, 9*(0), 129-144. http://dx.doi.org/10.1016/j.edurev.2013.01.002

Panadero, E., Jonsson, A., & Botella, J. (2017). Effects of self-assessment on self-regulated learning and self-efficacy: Four meta-analyses. *Educational Research Review, 22*, 74-98. doi:https://doi.org/10.1016/j.edurev.2017.08.004

Panadero, E., & Romero, M. (2014). To rubric or not to rubric? The effects of self-assessment on self-regulation, performance and self-efficacy. *Assessment in Education: Principles, Policy & Practice, 21*(2), 133-148. doi:10.1080/0969594X.2013.877872

Parkes, J. (2013). Reliability in classroom assessment. In J. H. McMillan (Ed.), *SAGE Handbook of Research on Classroom Assessment* (pp. 107-123). Los Angeles, CA, London, New Dehli, Singapore, Washington DC: SAGE.

Popham, W. J. (1994). The instructional consequences of criterion referenced clarity. *Educational Measurement: Issues and Practice, 13*(4), 15-18, 30. doi:10.1111/j.1745-3992.1994.tb00565.x

*Popham, J. (1997). What's wrong and what's right with rubrics. *Educational Leadership, 55*(2), 72-75.

Reddy, Y. M., & Andrade, H. (2010). A review of rubric use in higher education. *Assessment & Evaluation In Higher Education, 35*(4), 435-448. doi:10.1080/02602930902862859

Reynolds-Keefer, L. (2010). Rubric-referenced assessment in teacher preparation: An opportunity to learn by using. *Practical Assessment Research & Evaluation, 15*(8). Retrieved from http://pareonline.net/getvn.asp?v=15&n=8

*Rezaei, A. R., & Lovorn, M. G. (2010). Reliability and validity of rubrics for assessment through writing. *Assessing Writing, 15*(1), 18-39. doi:http://dx.doi.org/10.1016/j.asw.2010.01.003

Ritchie, S. M. (2016). Self-assessment of video-recorded presentations: Does it improve skills? *Active Learning in Higher Education, 17*(3) 207-221. doi: 10.1177/1469787416654807

Rochford, L., & Borchert, P. S. (2011). Assessing higher level learning: Developing rubrics for case analysis. *Journal of Education for Business, 86*, 258-265. doi: 10.1080/08832323.2010.512319

Rust, C., Price, M., & O'Donovan, B. (2003). Improving students' learning by developing their understanding of assessment criteria and processes. *Assessment & Evaluation In Higher Education, 28*(2), 147-164.

*Sadler, D. R. (2009). Indeterminacy in the use of preset criteria for assessment and grading. *Assessment & Evaluation in Higher Education, 34*(2), 159-179. doi: 10.1080/02602930801956059

*Sadler, D. R. (2014). The futility of attempting to codify academic achievement standards. *Higher Education, 67,* 273-288. doi: 10.1007/s10734-013-9649-1

Shay, S. B. (2004). The assessment of complex performance: A socially situated interpretive act. *Harvard Business Review 74*(3): 307-329

*Shipman, D., Roa, M., Hooten, J., & Wang, Z. J. (2012). Using the analytic rubric as an evaluation tool in nursing education: The positive and the negative. *Nurse Education Today, 32*(3), 246-249. doi:10.1016/j.nedt.2011.04.007

Soderstrom, N. C., & Bjork, R. A. (2015). Learning versus performance an integrative review. *Perspectives on Psychological Science, 10*(2), 176-199.

Stake, R. E. (2004). *Standards-based & responsive evaluation*. Thousand Oaks, CA: Sage.

Struyven, K., Dochy, F., & Janssens, S. (2005). Students' perceptions about evaluation and assessment in higher education: A review. *Assessment & Evaluation In Higher Education, 30*(4), 331-347.

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

Taylor, F. W. (1911). *The principles of scientific management*. New York: Harper & Brothers.

Tomas, C., Whitt, E., Lavelle-Hill, R., & Severn, K. (2019). Modeling holistic marks with analytic rubrics. Frontiers in Education: Assessment, Testing and Applied Measurement, 4(89). doi: 10.3389/feduc.2019.00089

*Torrance, H. (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. doi:10.1080/09695940701591867

*Torrance, H. (2012). Formative assessment at the crossroads: conformative, deformative and transformative assessment. *Oxford Review of Education, 38*(3), 323-342. doi: http://dx.doi.org/10.1080/03054985.2012.689693

Turley, E. D., & Gallagher, C. W. (2008). On the "uses" of rubrics: Reframing the great rubric debate. *The English Journal, 97*(4), 87-92. doi:10.2307/30047253

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

*Wilson, M. (2006). *Rethinking rubrics in writing assessment*. Porthsmouth, NH: Heinemann.

*Wilson, M. (2007). Why I won't be using rubrics to respond to students' writing. *The English Journal, 96*(4), 62-66. doi:10.2307/30047167

*Wollenschläger, M., Hattie, J., Machts, N., Möller, J., & Harms, U. (2016). What makes rubrics effective in teacher-feedback? Transparency of learning goals is not enough. *Contemporary Educational Psychology*. doi:http://dx.doi.org/10.1016/j.cedpsych.2015.11.003

Appendix 1. Rubric references from the combined authors' reference manager libraries

1. 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.
2. Andrade, H. (1999). *The role of instructional rubrics and self-assessment in learning to write: A smorgasbord of findings.* Paper presented at the Annual Meeting of the American Educational Research Association, Montreal, Canada.
3. Andrade, H. G. (1999). Student self-assessment: At the intersection of metacognition and authentic assessment. Paper Presented at the Annual Meeting of the American Educational Research Association.
4. Andrade, H. (2000). Using rubrics to promote thinking and learning. *Educational Leadership, 57*(5), 13-18.
5. Andrade, H. (2005). Teaching with rubrics: The good, the bad, and the ugly. *College Teaching, 53*(1), 27-32. doi:10.3200/CTCH.53.1.27-31
6. Andrade, H. (2007). Self-assessment through rubrics. *Educational Leadership, 65*(4), 60-63.
7. Andrade, H. (2010). Students as the definitive source of formative assessment: Academic self-assessment and the self-regulation of learning. In H. J. Andrade & G. J. Cizek (Eds.), *Handbook of formative assessment* (pp. 90-105). New York: Routledge.
8. Andrade, H. L. (2006). The trouble with a narrow view of rubrics. *English Journal*, 9-9.
9. Andrade, H., & Du, Y. (2005). Student perspectives on rubric-referenced assessment. *Practical Assessment, Research & Evaluation, 10*(3), 1-11. Retrieved from http://pareonline.net/getvn.asp?v=10&n=3
10. Andrade, H., & Du, Y. (2007). Student responses to criteria-referenced self-assessment. *Assessment & Evaluation In Higher Education, 32*(2), 159-181.
11. Andrade, H., & Valtcheva, A. (2009). Promoting learning and achievement through self-assessment. *Theory Into Practice, 48*(1), 12-19. doi:10.1080/00405840802577544

12. Andrade, H., Buff, C., Terry, J., Erano, M., & Paolino, S. (2009). Assessment-driven improvements in middle school students' writing. *Middle School Journal, 40*(4), 4-12.
13. Andrade, H., Du, Y., & Mycek, K. (2010). Rubric-referenced self-assessment and middle school students' writing. *Assessment in Education: Principles, Policy & Practice*, *17*(2), 199-214. doi:10.1080/09695941003696172
14. Andrade, H., Du, Y., & Wang, X. (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 Practices, 27*(2), 3-13.
15. Andrade, H., Wang, X. L., Du, Y., & Akawi, R. L. (2009). Rubric-referenced self-assessment and self-efficacy for writing. *Journal of Educational Research, 102*(4), 287-301.
16. Arter, J., & Chappuis. J. (2007). *Creating and recognizing quality rubrics.* Upper Saddle River, NJ: Pearson/Merrill Prentice Hall.
17. Arter, J., & McTighe, J. (2000). *Scoring rubrics in the classroom: Using performance criteria for assessing and improving student performance*: Corwin Press.
18. Aschbacher, P. R. (1999). *Developing indicators of classroom practice to monitor and support school reform. CSE Technical Report 513*. Los Angeles.
19. Ash, S. L., Clayton, P. H., & Atkinson, M. P. (2005). Integrating reflection and assessment to capture and improve student learning. *Michigan Journal of Community Service Learning, 11*(2), 49‑60.
20. Avanzino, S. (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.
21. Bacchus, R., Colvin, E., Knight, E. B., & Ritter, L. (2019). When rubrics aren't enough: Exploring exemplars and student rubric co-construction. *Journal of Curriculum and Pedagogy*, 1-14. doi:10.1080/15505170.2019.1627617
22. Baker, E. L. (1994). Learning-based assessments of history understanding. *Educational Psychologist, 29,* 97–106.

23. Balan, A. (2012). *Assessment for learning. A case study in mathematics education* (Doctoral dissertation). Malmö University, Malmö.
24. Baryla, E., Shelley, G., & Trainor, W. (2012). Transforming rubrics using factor analysis. *Practical Assessment Research & Evaluation, 17*(4), 1-7.
25. Bauer, C. F., & Cole, R. (2012). Validation of an assessment rubric via controlled modification of a classroom activity. *Journal of Chemical Education, 89*(9), 1104‑1108.
26. Bay, E., & Kotaman, H. (2011). Examination of the impact of rubric use on achievement in teacher education. *New Educational Review, 24*(2), 283-292.
27. Bearman, M., & Ajjawi, R. (2019). Can a rubric do more than be transparent? Invitation as a new metaphor for assessment criteria. *Studies in Higher Education*, 1-10. doi:10.1080/03075079.2019.1637842
28. Becker, A. (2016). Student-generated scoring rubrics: Examining their formative value for improving ESL students' writing performance. *Assessing Writing, 29*, 15-24. doi:http://dx.doi.org/10.1016/j.asw.2016.05.002
29. 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.
30. Bell, A., Mladenovic, R., & Price, M. (2013). Students' perceptions of the usefulness of marking guides, grade descriptors and annotated exemplars. *Assessment & Evaluation in Higher Education, 38*, 769–788.
31. Besterfield-Sacre, M., Gerchak, J., Lyons, M. R., Shuman, L. J., & Wolfe, H. (2004). Scoring concept maps: An integrated rubric for assessing engineering education. *Journal of Engineering Education, 93*(2), 105-115. doi:10.1002/j.2168-9830.2004.tb00795.x
32. Bissell, A. N., & Lemons, P. R. (2006). A new method for assessing critical thinking in the classroom. *BioScience*, *56*, 66–72.
33. Bolton, C.F. 2006. Rubrics and adult learners: Andragogy and assessment. *Assessment Update, 18*(3), 5–6.
34. 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, CA.

35. 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, CA.

36. Boulet, J. R, Rebbecchi, T.A., Denton, E.C., Mckinley, D., & Whelan, G. P. (2004). Assessing the written communication skills of medical school graduates. *Advances in Health Sciences Education, 9,* 47–60.

37. Bowen, T. (2017). Assessing visual literacy: a case study of developing a rubric for identifying and applying criteria to undergraduate student learning. *Teaching in Higher Education, 22,* 705–719.

38. Bresciani, M. J., Oakleaf, M., Kolkhorst, F., Nebeker, C., Barlow, J., Duncan, K., & Hickmott, J. (2009). Examining design and inter-rater reliability of a rubric measuring research quality across multiple disciplines. *Practical Assessment, Research & Evaluation, 14*(12), 1-7.

39. Britton, E., Simper, N., Leger, A., & Stephenson, J. (2017). Assessing teamwork in undergraduate education: a measurement tool to evaluate individual teamwork skills. *Assessment & Evaluation in Higher Education, 42,* 378–397.

40. 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.

41. Brookhart, S. M. (2013). *How to create and use rubrics for formative assessment and grading*. Virginia, USA: ASCD.

42. Brookhart, S. M. (2018). Appropriate criteria: Key to effective rubrics. *Frontiers in Education, 3*(22). doi:10.3389/feduc.2018.00022

43. Brookhart, S. M., & Chen, F. (2015). The quality and effectiveness of descriptive rubrics. *Educational Review, 67*(3), 343-368. doi:10.1080/00131911.2014.929565

44. 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*(2), 105-121. doi:10.1016/j.asw.2004.07.001

45. Bryant, C. L., Maarouf, S., Burcham, J., & Greer, D. (2016). The examination of a teacher candidate assessment rubric: A confirmatory factor analysis. *Teaching and*

Teacher Education, 57, 79-96. doi:http://dx.doi.org/10.1016/j.tate.2016.03.012

46. Campbell, A. (2005). Application of ICT and rubrics to the assessment process where professional judgment is involved: the features of an e-marking tool. *Assessment & Evaluation in Higher Education, 30*(5), 529–37.
47. Case, S. (2007). Reconfiguring and realigning the assessment feedback processes for an undergraduate criminology degree. *Assessment and Evaluation in Higher Education, 32*(3), 285-299.
48. Cebrián, M., & Monedero Moya, J. J. (2014). Evolution in the design and functionality of rubrics: from "square" rubrics to "federated" rubrics. *Revista de Docencia Universitaria, 12*(1), 81-98.
49. Cebrián, M., Serrano, J., & Ruiz, M. (2014). eRubrics in cooperative assessment of learning at university. *Comunicar, 43*. doi:10.3916/C43-2014-15
50. Chan, Z., & Ho, S. (2019). Good and bad practices in rubrics: The perspectives of students and educators. *Assessment & Evaluation In Higher Education, 44*(4), 533-545. doi:10.1080/02602938.2018.1522528
51. Chapman, V. G., & Inman, M. D. (2009). A conundrum: Rubrics or creativity/metacognitive development? *Educational Horizons*, 198-202.
52. Chasteen, S. V., Pepper, E. R., Caballero, M. D., Pollock, S. J., & Perkins, K. K. (2012). Colorado upper-division electrostatics diagnostic: A conceptual assessment for the junior level. *Physical Review Special Topics – Physics Education Research, 8*(2).
53. Cheng, M. W. T., & Chan, C. K. Y. (2019). An experimental test: Using rubrics for reflective writing to develop reflection. *Studies In Educational Evaluation, 61*, 176-182. doi:https://doi.org/10.1016/j.stueduc.2019.04.001
54. Chi, E. (2001). Comparing holistic and analytic scoring for performance assessment with many-facet rasch model. *Journal of Applied Measurement, 2*, 379–388.
55. 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.
56. 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.

57. Ciorba, C. R., & Smith, N. Y. (2009). Measurement of instrumental and vocal undergraduate performance juries using a multidimensional assessment rubric. *Journal of Research in Music Education, 57*(1), 5‑15.

58. 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, CA.

59. Cockett, A., & Jackson, C. (2018). The use of assessment rubrics to enhance feedback in higher education: An integrative literature review. *Nurse Education Today, 69*, 8-13. doi:https://doi.org/10.1016/j.nedt.2018.06.022

60. Coe, M., Hanita, M., Nishioka, V., & Smiley, R. (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.

61. Cohen, E. G., Lotan, R. A., Abram, P. L., Scarloss, B. A. & Schultz, S. E. (2002). Can groups learn? *Teachers College Record, 104*(6), 1045–1068.

62. Company, P., Contero, M., Otey, J., Camba, J. D., Agost, M. J., & Pérez-López, D. (2017). Web-Based System for Adaptable Rubrics: Case Study on CAD Assessment. *Educational Technology & Society, 20*(3), 24-41.

63. Cooper, B. S., & Gargan, A. (2009). Rubrics in education. *Phi Delta Kappan, 91*(1), 54-55.

64. Dandis, M. A. I. (2013). *Teachers' beliefs about assessment for learning: Introducing rubrics in Secondary Education.* (Ph.D.), Universidad de Granada.

65. Davidowitz, B., Rollnick, M., & Fakudze, C. (2005). Development and application of a rubric for analysis of novice students' laboratory flow diagrams. *International Journal of Science Education, 27*, 43–59.

66. Dawson, P. (2017). Assessment rubrics: towards clearer and more replicable design, research and practice. *Assessment & Evaluation in Higher Education, 42*, 347-360. doi:10.1080/02602938.2015.1111294

67. De Wever, B., Van Keer, H., Schellens, T., & Valcke, M. (2011). Assessing collaboration in a Wiki: The reliability of university students' peer assessment. *The*

Internet and Higher Education, 14(4), 201‑206.

68. 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. Paper presented at the annual meeting of the American Association of Colleges for Teacher Education.
69. DeRemer, M. L. (1998). Writing assessment: Raters' elaboration of the rating task. *Assessing Writing, 5,* 7–29.
70. Dinur, A., & Sherman, H. (2009). Incorporating outcomes assessment and rubrics into case instruction. *Journal of Behavioral and Applied Management, 10*(2), 291‑311.
71. Docktor, J., & Heller, K. (2009). Assessment of student problem solving processes. *AIP Conference Proceedings, 1179*: 133‑136.
72. 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. University of Oklahoma, Norman, OK.
73. 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.
74. Dyson, Ben, Placek, J. H., Graber, K. C., Fisette, J. L., Rink, J., Zhu, W., & Avery, M. (2011). Development of PE metrics elementary assessments for national physical education standard 1. *Measurement in Physical Education and Exercise Science, 15*(2), 100‑118.
75. Facione, N. C., & Facione, P. A. (1996). Externalizing the critical thinking in knowledge development and clinical judgment. *Nursing Outlook, 44,* 129–136.
76. Fernández Sánchez, M. J., Lucero, M., & Montanero, M. (2014). Redacción de textos narrativos en educación primaria: Comparativa de recursos didácticos para su evaluación. *Cadernos de Linguagem e Sociedade, 15*(2), 79-97.
77. Finson, K. D., & Ormsbee, C. K. (1998). Rubrics and their use in inclusive science. *Intervention in School and Clinic, 34*(2), 79-88. doi:10.1177/105345129803400203

78. Fleenor, J. W., Fleenor, J. B., & Grossnickle, W. F. (1996). Interrater reliability and agreement of performance ratings: A methodological comparison. *Journal of Business and Psychology, 10*, 367–80.

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

80. 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.

81. Fraile, J., Panadero, E., & Pardo, R. (2017). Co-creating rubrics: useful or waste of time? The effects of establishing assessment criteria with students on self- regulation, self-efficacy and performance. *Studies In Educational Evaluation, 53*, 69-76. doi:10.1016/j.stueduc.2017.03.003

82. Fraile, J., Pardo, R., & Panadero, E. (2017). ¿Cómo emplear las rúbricas para implementar una verdadera evaluación formativa? *Revista Complutense de Educación, 28*(4), 1321-1334. doi:http://dx.doi.org/10.5209/RCED.51915

83. Fraser, L., Harich, K., Norby, J., Brzovic, K., Rizkallah, T., & Loewy, D. (2005). Diagnostic and value-added assessment of business writing. *Business Communication Quarterly, 68*(3), 290‑305.

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

85. Gantt, L. T. (2010). Using the Clark simulation evaluation rubric with associate degree and baccalaureate nursing students. *Nursing Education Perspectives, 31*(2), 101-105.

86. Garcia-Ros, R. (2011). Análisis y validación de una rúbrica para evaluar habilidades de presentación oral en contextos universitarios. *Electronic Journal of Research in Educational Psychology*, *9*(3).

87. García-ros, R. (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.

88. 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.

89. Gerretson, H., & Golson, E. (2005). Synopsis of the use of course-embedded assessment in a medium sized public university's general education program. *The Journal of General Education, 54*(2), 139‑149.
90. 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.
91. Goodrich Andrade, H. (1996). Understanding rubrics. *Educational Leadership*.
92. Goodrich Andrade, H. (2000). Using rubrics to promote thinking and learning. *Educational Leadership, 57*(5).
93. Goodrich Andrade, H. (2001). The effects of instructional rubrics on learning to write. *Current Issues in Education, 4*(4). Retrieved from http://cie.ed.asu.edu/volume4/number4/
94. Goodrich Andrade, H., & Boulay, B. A. (2003). Role of rubric-referenced self-assessment in learning to write. *The Journal of Educational Research, 97*(1), 21-34. doi:10.1080/00220670309596625
95. Graham, M., Milanowski, A., & Miller, J. (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.
96. Green, R., & Bowser, M. (2006). Observations from the field: Sharing a literature review rubric. *Journal of Library Administration, 45*(1), 185 - 202. doi:10.1300/J111v45n01_10
97. Gregait, L. H., Johnsen, D. R., & Nielsen, P. S. (1997). *Improving evaluation of student participation in physical education through self-assessment*. Retrieved from
98. Hack, C. (2015). Analytical rubrics in higher education: A repository of empirical data. *British Journal of Educational Technology, 46*(5), 924-927. doi:10.1111/bjet.12304
99. Hafner, O. C., & Hafner, P. (2003). Quantitative analysis of the rubric as an assessment tool: An empirical study of student peer-group rating. *International Journal of Science Education, 25*(12), 1509-1528. doi:10.1080/0950069022000038268

100. Halonen, J. S., Bosack, T., Clay, S., McCarthy, M., Dunn, D. S., Hill Iv, G. W., . . . Whitlock, K. (2003). A rubric for learning, teaching, and assessing scientific inquiry in psychology. *Teaching of Psychology, 30*(3), 196-208. doi:10.1207/s15328023top3003_01

101. Hancock, A. B., & Brundage, S. B. (2010). Formative feedback, rubrics, and assessment of professional competency through a speech-language pathology graduate program. *Journal of Epidemiology and Community Health, 39*, 110–119.

102. Hanna, M. A., & Smith., J. (1998). Using rubrics for documentation of clinical work supervision. *Counselor Education and Supervision, 37*(4), 269-278.

103. Harnden, J. (2005). Scientific inquiry scoring guides. *Science Scope, 28*(4), 52‑54.

104. Hay, P. J., & Macdonald, D. (2008). (Mis)Appropriations of criteria and standards-referenced assessment in a performance-based subject. *Assessment in Education: Principles, Policy & Practice, 15*(2), 153‑168.

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

106. 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, GA.

107. Hitt, A. M., & Helms, E. C. (2009). Best in show: Teaching old dogs to use new rubrics. *The Professional Educator, 33*(1), 1.

108. Holmstedt, P., Jonsson, A., & Aspelin, J. (2018). Learning to see new things: Using criteria to support pre-service teachers' discernment in the context of teachers' relational work. *Frontiers in Education: Assessment, Testing and Applied Measurement, 3*(54).

109. Howell, R. 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*, 31–49.

110. Howell, R. J. (2014). Grading rubrics: hoopla or help? *Innovations in Education and*

Teaching International, 51, 400-410.

111. Hudson, J., Bloxham, S., den Outer, B., & Price, M. (2017). Conceptual acrobatics: talking about assessment standards in the transparency era. *Studies in Higher Education, 42*(7), 1309-1323. doi:10.1080/03075079.2015.1092130

112. Humphry, S. M., & Heldsinger, S. A. (2014). Common structural design features of rubrics may represent a threat to validity. *Educational Researcher, 43*(5), 253-263. doi:10.3102/0013189x14542154

113. Iacobucci, D., & Duhachek, A. (2003). Advancing alpha: Measuring reliability with confidence. *Journal of Consumer Psychology, 13*(4), 478‑487.

114. Isaacson, J. J., & Stacy, A. S. (2009). Rubrics for clinical evaluation: Objectifying the subjective experience. *Nurse Education in Practice, 9*(2), 134-140.

115. Ito, H. (2015). Is a rubric worth the time and effort? Conditions for its success. *International Journal of Learning, Teaching and Educational Research, 10*(2).

116. Jeong, H. (2015). What is your teacher rubric? Extracting teachers' assessment constructs. *Practical Assessment Research & Evaluation, 20*, 1-13. Retrieved from http://pareonline.net/getvn.asp?v=20&n=6

117. Jescovitch, L. N., Scott, E. E., Cerchiara, J. A., Doherty, J. H., Wenderoth, M. P., Merrill, J. E., . . . Haudek, K. C. (2019). Deconstruction of Holistic Rubrics into Analytic Rubrics for Large-Scale Assessments of Students' Reasoning of Complex Science Concepts. *Practical Assessment, Research & Evaluation, 24*(7), 2.

118. 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.

119. 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*(2124-138).

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

121. Jones, L., Allen, B., Dunn, P., & Brooker, L. (2016). Demystifying the rubric: A five-step pedagogy to improve student understanding and utilisation of marking criteria. *Higher Education Research & Development*, 1-14.

doi:10.1080/07294360.2016.1177000

122. Jonsson, A. (2010). The use of transparency in the "Interactive Examination" for student teachers. *Assessment in Education: Principles, Policy & Practice, 17*(2), 183-197.

123. Jonsson, A. (2013). *Communicating expectations through the use of rubrics*. Paper presented at the EARLI Conference 2013, Munich, Germany.

124. Jonsson, A. (2014). Rubrics as a way of providing transparency in assessment. *Assessment & Evaluation In Higher Education*, 1-13. doi:10.1080/02602938.2013.875117

125. Jonsson, A., & Panadero, E. (2016). The use and design of rubrics to support Assessment for Learning In D. R. Carless (Ed.), *Scaling up Assessment for Learning*.

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

127. Joseph, S., Rickett, C., Northcote, M., & Christian, B. J. (2019). 'Who are you to judge my writing?': Student collaboration in the co-construction of assessment rubrics. *New Writing*, 1-19. doi:10.1080/14790726.2019.1566368

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

129. Kenworthy, A. L., & Hrivnak, G. A. (2014). To Rubric or Not to Rubric: That Is the Question. *Journal of Management Education, 38*(3), 345-351. doi:10.1177/1052562914530103

130. Kerby, D., & Romine, J. (2010). Develop oral presentation skills through accounting curriculum design and course-embedded assessment. *Journal of Education for Business, 85*, 172–179.

131. Kilgour, P., Northcote, M., Williams, A., & Kilgour, A. (2019). A plan for the co-construction and collaborative use of rubrics for student learning. *Assessment & Evaluation In Higher Education*, 1-14. doi:10.1080/02602938.2019.1614523

132. Knight, L. A. (2006). Using rubrics to assess information literacy. *Reference Services Review, 34*, 43–55.

133. Knoch, U. (2009). Diagnostic assessment of writing: A comparison of two rating scales. *Language Testing, 26*(2), 275-304. doi:10.1177/0265532208101008

134. Kocakulah, M. S. (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

135. Kohn, A. (2006). The trouble with rubrics. *English Journal, 95*(4), 12-15.

136. 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.

137. Kutlu, O., Yıldırım, O., & Bilican, S. (2010). The comparison of the views of teachers with positive and negative attitudes towards rubrics. *Procedia - Social and Behavioral Sciences, 9*(0), 1566-1573. doi:http://dx.doi.org/10.1016/j.sbspro.2010.12.366

138. Lapsley, R., & Moody, R. (2007). Teaching tip: Structuring a rubric for online course discussions to assess both traditional and non-traditional students. *Journal of American Aacademy of Business, 12(*1), 167–172.

139. Lasater, K. (2007). Clinical judgment development: Using simulation to create an assessment rubric. *Journal of Nursing Education, 46*(11).

140. Latifa, A., Rahman, A., Hamra, A., Jabu, B., & Nur, R. (2015). Developing a practical rating rubric of speaking test for university students of English in Parepare, Indonesia. *English Language Teaching, 8*, 166–177.

141. Laveault, D., & Miles, C. (2002). *The study of individual differences in the utility and validity of rubrics in the learning of writing ability*.

142. Lee, E., & Sohyun L. (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.

143. Lewis, L. K., Stiller, K., & Hardy, F. (2008). A clinical assessment tool used for physiotherapy students - is it reliable? *Physiotherapy Theory and Practice, 24*(2), 121‑134.

144. Lindblom-Ylänne, S., Pihlajamäki, H., & Kotkas, T. (2006). Self-, peer- and teacher-assessment of student essays. *Active Learning in Higher Education, 7*(1), 51-62. doi:10.1177/1469787406061148

145. Lipnevich, A. A., McCallen, L. N., Miles, K. P., & Smith, J. K. (2014). Mind the

gap! Students' use of exemplars and detailed rubrics as formative assessment. *Instructional Science, 42*(4), 539-559. doi:10.1007/s11251-013-9299-9

146. Loeffler, K. A. (2005). No more Friday spelling tests? An alternative spelling assessment for students with learning disabilities. *Teaching Exceptional Children, 37*(4), 24‑27.

147. Lovorn, M. G., & Rezaei, A. R. (2011). Assessing the assessment: Rubrics training for pre-service and new in-service teacher. *Practical Assessment, Research & Evaluation, 16*(16).

148. Lucander, H., Knutsson, K., Salé, H., & Jonsson, A. (2012). "I'll never forget this": Evaluating a pilot workshop in effective communication. *Journal of Dental Education, 76*(10), 1311-1316.

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

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

151. Mabry, L. (1999). Writing to the rubric: Lingering effects of traditional standardized testing on direct writing assessment. *Phi Delta Kappan, 80*(9), 673-679.

152. Mansilla, V. B., Duraisingh, E. D., Wolfe, C. R., & Haynes, C. (2009). Targeted assessment rubric: An empirically grounded rubric for interdisciplinary writing. *Journal of Higher Education, 80*(3), 334‑353.

153. Marín-García, J. A., & Santandreu-Mascarell, C. (2015). ¿Qué sabemos sobre el uso de rúbricas en la evaluación de asignaturas universitarias? *Intangible Capital*. doi:http://dx.doi.org/10.3926/ic.538

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

155. Mason, C. Y., Steedly, K. M., & Thormann, M. S. (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.

156. 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, CA.

157. McCormick, M. J., Dooley, K. E., Lindner, J. R., & Cummins, R. L. (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.

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

159. Meier, S. L., Rich, B. S., & Cady, J. A. (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.

160. Menéndez-Varela, J. L., & Gregori-Giralt, E. (2017). The reliability and sources of error of using rubrics-based assessment for student projects. *Assessment & Evaluation In Higher Education*, 1-12. doi:10.1080/02602938.2017.1360838

161. Mertler, C. A. (2001). Designing scoring rubrics for your classroom. *Practical Assessment Research & Evaluation, 7*(25). Retrieved from http://PAREonline.net/getvn.asp?v=7&n=25

162. Moni, R. W., & Moni, K. B. (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

163. Moni, R. W., Beswick, E., & Moni, K. B. (2005). Using student feedback to construct an assessment rubric for a concept map in physiology. *Advances in Physiology Education, 29*(4), 197-203. doi:10.1152/advan.00066.2004

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

165. Moskal, B. M. (2000). Scoring rubrics: What, when and how? *Practical Assessment Research & Evaluation, 7*(3). Retrieved from

http://PAREonline.net/getvn.asp?v=7&n=3

166. Moskal, B. M. (2003). Recommendations for developing classroom performance assessments and scoring rubrics. *Practical Assessment Research & Evaluation, 8*(14). Retrieved from http://PAREonline.net/getvn.asp?v=8&n=14

167. Moskal, B. M., & Leydens, J. A. (2000). Scoring rubric development: Validity and reliability. *Practical Assessment, Research & Evaluation, 7*(10). Retrieved from http://PAREonline.net/getvn.asp?v=7&n=10

168. 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*.

169. Mott, M. S., Chessin, D. A., Sumrall, W. J., Rutherford, A. S., & Moore, V. J. (2011). Assessing student scientific expression using media: the mediaenhanced science presentation rubric (MESPR). *Journal of STEM Education: Innovations and Research, 12*(1&2), 33‑41.

170. Mullen, Y. K. (2003). *Student Improvement in Middle School Science. Master's thesis.*, University of Wisconsin.

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

172. Newman, L. R., Lown, B. A., Jones, R. N., Johansson, A., & Schwartzstein, R. M. (2009). Developing a peer assessment of lecturing instrument: Lessons learned. *Journal of the Association of American Medical Colleges, 84*(8), 1104‑1110.

173. Nicholson, P., Gillis, S., & Dunning, A. M. (2009). The use of scoring rubrics to determine clinical performance in the operating suite. *Nurse Education Today, 29*(1), 73‑82.

174. Norcini, J. J. (1999). Standards and reliability in evaluation: When rules of thumb don't apply. *Academic Medicine, 74*(10), 1088‑1090.

175. Nordrum, L., Evans, K., & Gustafsson, M. (2013). Comparing student learning experiences of in-text commentary and rubric-articulated feedback: Strategies for formative assessment. *Assessment & Evaluation In Higher Education, 38*(8), 919-

940. doi:10.1080/02602938.2012.758229

176. O'Donovan, B., Price, M., & Rust, C. (2004). Know what I mean? Enhancing student understanding of assessment standards and criteria. *Teaching in Higher Education, 9,* 325-335.

177. Oakleaf, M. J. (2006). Assessing information literacy skills: A rubric approach. (Doctoral dissertation). University of North Carolina, Chapel Hill, NC. UMI No. 3207346.

178. O'Brien, C. E., Franks, A. M., & Stowe, C. D. (2008). Multiple rubric-based assessments of student case presentations. *American Journal of Pharmaceutical Education, 72*(3), Article 58.

179. 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.

180. Orsmond, P., & Merry, S. (1996). The importance of marking criteria in the use of peer assessment. *Assessment & Evaluation in Higher Education, 21*(3), 239-250.

181. Orsmond, P., Merry, S., & Reiling, K. (1997). A study in self-assessment: Tutor and students' perceptions of performance criteria. *Assessment & Evaluation in Higher Education, 22*(4), 357–368.

182. 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.

183. Pagano, N., Bernhardt, S. A., Reynolds, D., Williams, M., & McCurrie, M. K. (2008). An inter-institutional model for college writing assessment. *College Composition and Communication, 60*(2), 285‑320.

184. Panadero, E. (2011). Instructional help for self-assessment and self-regulation: Evaluation of the efficacy of self-assessment scripts vs. rubrics. Doctoral dissertation. Universidad Autónoma de Madrid, Madrid, Spain.

185. Panadero, E., & Alonso-Tapia, J. (2011). El papel de las rúbricas en la autoevaluación y autorregulación del aprendizaje. In K. Bujan, I. Rekalde, & P. Aramendi (Eds.), *La Evaluación de Competencias en la Educación Superior. Las rúbricas como instrumento de evaluación* (pp. 97-111). Madrid: Eduforma.

186. Panadero, E., & Alonso-Tapia, J. (2012). *Efectos de rúbricas y guiones de autoevaluación en la autorregulación, aprendizaje, auto-eficacia y motivación en estudiantes de educación secundaria.* Paper presented at the Encuentro Jóvenes Investigadores en Motivación y Emoción, Universidad Autónoma de Madrid (Spain).

187. Panadero, E., & Jonsson, A. (2013). The use of scoring rubrics for formative assessment purposes revisited: A review. *Educational Research Review, 9*(0), 129-144. doi:http://dx.doi.org/10.1016/j.edurev.2013.01.002

188. Panadero, E., & Romero, M. (2012). *Uso de las rúbricas evaluación para fomentar el aprendizaje autorregulado/autónomo.* Paper presented at the CIDUI 2012, Barcelona.

189. Panadero, E., & Romero, M. (2014). To rubric or not to rubric? The effects of self-assessment on self-regulation, performance and self-efficacy. *Assessment in Education: Principles, Policy & Practice, 21*(2), 133-148. doi:10.1080/0969594X.2013.877872

190. Panadero, E., Alonso-Tapia, J., & Huertas, J. A. (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

191. Panadero, E., Alonso-Tapia, J., & Huertas, J. A. (2014). Rubrics vs. self-assessment scripts: Effects on first year university students' self-regulation and performance. *Infancia y Aprendizaje: Journal for the Study of Education and Development, 37*(1), 149-183. doi:10.1080/02103702.2014.881655

192. Panadero, E., Alonso-Tapia, J., & Reche, E. (2013). Rubrics vs. self-assessment scripts effect on self-regulation, performance and self-efficacy in pre-service teachers. *Studies In Educational Evaluation, 39*(3), 125-132. doi:10.1016/j.stueduc.2013.04.001

193. Panadero, E., Jonsson, A., & Botella, J. (2017). Effects of self-assessment on self-regulated learning and self-efficacy: Four meta-analyses. *Educational Research Review, 22*, 74-98.

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

195. Pascual-Gómez, I., Lorenzo-Llamas, E. M., & Monge-López, C. (2015). Análisis de validez en la evaluación entre iguales: un estudio en educación superior. *RELIEVE-Revista Electrónica de Investigación y Evaluación Educativa, 21*(1).

196. Peach, B. E., Mukherjee, A., & Hornyak, M. (2007). Assessing critical thinking: A college's journey and lessons learned. *Journal of Education for Business, 82*(6), 313‑320.

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

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

199. Perchemlides, N., & Coutant, C. (2004). Growing beyond grades. *Educational Leadership, 62*(2), 53-56.

200. Petkov, D., & Petkova, O. (2006). Development of scoring rubrics for IS projects as an assessment tool. *Issues in Informing Science and Information Technology, 3*, 499-510.

201. 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.

202. Piscitello, M. E. (2001). *Using rubrics for assessment and evaluation in art.* Saint Xavier University, Chicago, IL.

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

204. Popham, J. (1997). What's wrong and what's right with rubrics. *Educational Leadership, 55*(2), 72-75.

205. 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. Paper Presented at Annual Meeting of the American Educational Research Association.

206. Prins, F. J., de Kleijn, R., & van Tartwijk, J. (2015). Students' use of a rubric for research theses. *Assessment & Evaluation In Higher Education*, 1-23.

doi:10.1080/02602938.2015.1085954

207. 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.

208. Reddy, M. 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.

209. Reddy, Y. M. (2007). Effects of rubrics on enhancement of student learning. *Educate, 7*(1), 3-17.

210. Reddy, Y. M., & Andrade, H. (2010). A review of rubric use in higher education. *Assessment & Evaluation In Higher Education, 35*(4), 435-448. doi:10.1080/02602930902862859

211. Reitmeier, C. A., & Vrchota, D. A. (2009). Self-assessment of oral communication presentations in food science and nutrition. *Journal of Food Science Education, 8*(4), 88-92. doi:10.1111/j.1541-4329.2009.00080.x

212. Reitmeier, C. A., Svendsen, L. K., & Vrchota, D. A. (2004). Improving oral communication skills of students in food science courses. *Journal of Food Science Education, 3*(2), 15-20.

213. Reynolds-Keefer, L. (2010). Rubric-referenced assessment in teacher preparation: An opportunity to learn by using. *Practical Assessment Research & Evaluation, 15*(8). Retrieved from http://pareonline.net/getvn.asp?v=15&n=8

214. Rezaei, A. R., & Lovorn, M. G. (2010). Reliability and validity of rubrics for assessment through writing. *Assessing Writing, 15*(1), 18-39. doi:http://dx.doi.org/10.1016/j.asw.2010.01.003

215. Reznitskaya, A., Kuo, L-J., Glina, M., & Anderson, R. C. (2009). Measuring argumentative reasoning: What's behind the numbers? *Learning and Individual Differences, 19*(2), 219‑224.

216. Ritchie, S. M. (2016). Self-assessment of video-recorded presentations: does it improve skills? *Active Learning in Higher Education, 17*, 207-221.

217. 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.

218. Rochford, L., & Borchert, P. S. (2011). Assessing higher level learning: Developing rubrics for case analysis. *Journal of Education for Business, 86*(5), 258‑265.

219. Ross-Fisher, R. L. (2005). Developing effective success rubrics. *Kappa Delta Pi Record, 41*(3), 131-135.

220. Rusman, E., & Dirkx, K. (2017). Developing rubrics to assess complex (generic) skills in the classroom: How to distinguish skills' mastery levels? *Practical Assessment Research & Evaluation, 22*(12). Retrieved from http://pareonline.net/getvn.asp?v=22&n=12

221. Saddler, B., & Andrade, H. (2004). The writing rubric. *Educational Leadership, 62*(2), 48-52.

222. Sadler, D. R. (2007). Perils in the meticulous specification of goals and assessment criteria. *Assessment in Education: Principles, Policy & Practice, 14*(3), 387-392. doi:10.1080/09695940701592097

223. Sadler, D. R. (2009). Indeterminacy in the use of preset criteria for assessment and grading. *Assessment and Evaluation in Higher Education, 34*(2), 159-179.

224. Sadler, D. R. (2014). The futility of attempting to codify academic achievement standards. *Higher Education, 67*, 273–288.

225. Sadler, P. M., & Good, E. (2006). The impact of self- and peer-grading on student learning. *Educational Assessment, 11*(1), 1 - 31. doi:10.1207/s15326977ea1101_1

226. Sáiz-Manzanares, M. C., Cuesta Segura, I. I., Alegre Calderon, J. M., & Peñacoba Antona, L. (2017). Effects of different types of rubric-based feedback on learning outcomes. *Frontiers in Education, 2*(34). doi:10.3389/feduc.2017.00034

227. Sanchez, C. E., Atkinson, K. M., Koenka, A. C., Moshontz, H., & Cooper, H. (2017). Self-grading and peer-grading for formative and summative assessments in 3rd through 12th grade classrooms: A meta-analysis. *Journal of Educational Psychology*, No Pagination Specified. doi:10.1037/edu0000190

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

229. 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*(2), 151-170.

230. 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. *The Journal of General Education, 55*(2), 103-137. doi:10.1353/jge.2006.0025

231. 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. Nova Southeastern University, FL.

232. 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.

233. Schlitz, S. A., O'Connor M., Pang, Y., Stryker, D., Markell, S., Krupp, E., Byers, C., Jones, S. D., & Redfern, A. K. (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.

234. Schreiber, L. M., Paul, G. D., & Shibley, L. R. (2012). The development and test of the public speaking competence rubric. *Communication Education, 61*(3), 205‑233.

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

236. Shipman, D., Roa, M., Hooten, J., & Wang, Z. J. (2012). Using the analytic rubric as an evaluation tool in nursing education: The positive and the negative. *Nurse Education Today, 32*(3), 246-249. doi:10.1016/j.nedt.2011.04.007

237. Siegel, M. A., Halverson, K., Freyermuth, S., & Clark, C. G. (2011). Beyond grading: A series of rubrics for science learning in high school biology courses. *Science Teacher, 78*(1), 28‑33.

238. Silvestri, L., & Oescher, J. (2006). Using rubrics to increase the reliability of assessment in health classes. *International Electronic Journal of Health Education, 9*, 25-30.

239. Simon, M., & Forgette-Giroux, R. (2001). A rubric for scoring postsecondary academic skills. *Practical Assessment Research & Evaluation, 7*(18). Retrieved from http://PAREonline.net/getvn.asp?v=7&n=18

240. Smit, R., & Birri, T. Assuring the quality of standards-oriented classroom assessment with rubrics for complex competencies. *Studies In Educational Evaluation*(0).

doi:http://dx.doi.org/10.1016/j.stueduc.2014.02.002

241. Smit, R., Bachmann, P., Blum, V., Birri, T., & Hess, K. (2017). Effects of a rubric for mathematical reasoning on teaching and learning in primary school. *Instructional Science*, 1-20. doi:10.1007/s11251-017-9416-2

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

243. Spandel, V. (2006). In defense of rubrics. *English Journal, 96*(1).

244. Spence, L. K. (2010). Discerning writing assessment: Insights into an analytical rubric. *Language Arts, 87*(5), 337‑352.

245. Stellmack, M. A., Konheim-Kalkstein, Y. L., Manor, J. E., Massey, A. R., & Schmitz, J. A. P. (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

246. Stevens, D., & Levi, A. J. (2005). *Introduction to rubrics. An assessment tool to save grading time, convey effective feedback, and promote student learning*. Sterling, Virginia (USA): Stylus.

247. Stiggins, R. J., & Conklin, N. F. (1992). *In teachers' hands: Investigating the practices of classroom assessment*. Albany: SUNY Press.

248. 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.

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

250. 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.

251. Thaler, N., Kazemi, E., & Huscher, C. (2009). Developing a rubric to assess student learning outcomes using a class assignment. *Teaching of Psychology, 36*(2), 113-116. doi:10.1080/00986280902739305

252. Tierney, R., & Simon, M. (2004). What's still wrong with rubrics: Focusing on the

consistency of performance criteria across scale levels. *Practical Assessment Research & Evaluation, 9*(2). Retrieved from http://PAREonline.net/getvn.asp?v=9&n=2

253. Timmerman, B., Crotwell, E., Strickland, D. C., Johnson, R. L., & Payne, J. R. (2011). Development of a “universal” rubric for assessing undergraduates’ scientific reasoning skills using scientific writing. *Assessment & Evaluation in Higher Education, 36*(5), 509‑547.

254. Torrance, H. (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. doi:10.1080/09695940701591867

255. Torrance, H. (2012). Formative assessment at the crossroads: conformative, deformative and transformative assessment. *Oxford Review of Education, 38,* 323-342.

256. 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*(2), 264-286. doi:10.1002/sce.10004

257. Tractenberg, R. E., Umans, J. G., & McCarter, R. J. (2010). A mastery rubric: Guiding curriculum design, admissions and development of course objectives. *Assessment & Evaluation In Higher Education, 35*(1), 17-35.

258. Turley, E. D., & Gallagher, C. W. (2008). On the "uses" of rubrics: Reframing the great rubric debate. *The English Journal, 97*(4), 87-92. doi:10.2307/30047253

259. Urios, M. I., Rangel, E. R., Tomàs, R. B., Salvador, J. T., Garciá, F. C., & Piquer, C. F. (2015). Generic skills development and learning/assessment process: use of rubrics and student validation. *Journal of Science Education and Technology, 5,* 107–121.

260. Vandenberg, A., Stollak, M., Mckeag, L., & Obermann, D. (2010). GPS in the classroom: Using rubrics to increase student achievement. *Research in Higher Education Journal, 9,* 1‑10.

261. Various. (2014). Evaluación formativa mediante Erúbricas. *Revista de Docencia Universitaria, 12*(1).

262. Wald, H. S., Borkan, J. M., Taylor, J. S., Anthony, D., & Reis, S. P. (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.

263. Wallace, C. S., Prather, E. E., & Duncan, D. K. (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*.

264. 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, CA.

265. Wang, W. (2016). Using rubrics in student self-assessment: student perceptions in the English as a foreign language writing context. *Assessment & Evaluation In Higher Education*, 1-13. doi:10.1080/02602938.2016.1261993

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

267. 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, IL.

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

269. Wenzlaff, T. L., Fager, J. J., & Coleman, M. J. (1999). What is a rubric? Do practitioners and the literature agree? *Contemporary education, 70*(4), 41.

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

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

272. Wilson, M. (2006). *Rethinking rubrics in writing assessment*. Porthsmouth, NH: Heinemann.

273. Wilson, M. (2007). Why I won't be using rubrics to respond to students' writing. *The English Journal, 96*(4), 62-66. doi:10.2307/30047167

274. Wolf, K., & Stevens, E. (2007). The role of rubrics in advancing and assessing

student learning. *Journal of Effective Teaching, 7*(1), 3-14.

275. Wollenschläger, M., Hattie, J., Machts, N., Möller, J., & Harms, U. (2016). What makes rubrics effective in teacher-feedback? Transparency of learning goals is not enough. *Contemporary Educational Psychology*. doi:http://dx.doi.org/10.1016/j.cedpsych.2015.11.003

276. Wyatt-Smith, C., & Klenowski, V. (2013). Explicit, latent and meta-criteria: types of criteria at play in professional judgement practice. *Assessment in Education: Principles, Policy & Practice, 20*, 35-52.

277. Wylie, C., & Lyon, C. (2013). *Using the Formative Assessment Rubrics, Reflection and Observation Tools to Support Professional Reflection on Practice*: Formative Assessment for Teachers and Students.

278. Yopp, B. D., & Rehberger, R. (2009). A curriculum focus intervention's effects on prealgebra achievement. *Journal of Developmental Education, 33*(2), 28-30.

279. Young, C. (2013). Initiating self-assessment strategies in novice physiotherapy students: a method case study. *Assessment & Evaluation in Higher Education, 38*, 998–1011.

280. Zimmaro, D. H. (2004). Developing grading rubrics. Retrieved from http://www.utexas.edu/academic/mec/research/pdf/rubricshandout.pdf

Appendix 2. Excerpts extracted organized by publication including empirical evidence for the claim

Reference	Theme	Content	Empirical evidence for the claim
Popham, 1997	**Standardization and narrow the curriculum**	1. But what if the evaluative criteria in a rubric are linked only to the specific elements in a particular performance test?...Perhaps the commercial test publishers are eager to install task-specific evaluative criteria because such criteria permit more rapid scoring with a much greater likelihood of between-scorer agreement. But such criteria, from an instructional perspective, are essentially worthless. Teachers need evaluative criteria that capture the essential ingredients of the skill being measured, not the particular display of that skill applied to a specific task. (p. 73)	Anecdotal evidence. Popham's evidence seems to come from his direct observation of practices as an expert in the field of assessment. The paper is informational and based on his own experience, not an empirical study.
		2. Many rubrics now being billed as instructionally useful provide teachers and students with absolutely no cues about what is genuinely significant in a student's response, and they offer teachers no guidance on the key features of the tested skill. (p. 73)	
		3. Another shortcoming in many rubrics is excessive length: Busy teachers won't have anything to do with them. If we want rubrics to make a difference in classroom instruction, we need to create rubrics that teachers will use. Lengthy, overly detailed rubrics are apt to be used only by inordinately compulsive teachers. Many of the rubrics being circulated these days are lengthy and laden with details. After all, most of the earliest rubrics were created for use in large-scale, high-stakes assessments. If a state's high school diploma were to be based on how well a student functioned on an important statewide performance test—a writing sample, for instance—the architects of the accompanying rubric understandably might have leaned toward detailed scoring rules. In general, the more detailed and constraining a rubric's scoring rules, the greater the likelihood of between-rater agreement. For high-stakes tests, detailed rubrics were common. When educators and textbook publishers introduced rubrics for classroom use, many models came from these earlier large-scale assessments. (p. 74)	
		4. This problem stems less from rubrics themselves than from an error made by rubric users. A particularly prevalent misunderstanding occurs when rubric users become so caught up with the particulars of a given performance test that they begin thinking of the test as the skill itself. (p. 74)	
Mabry, 1999	**Standardization and narrow the curriculum**	5. As part of the larger argument that traditional psychometric concepts and practices undermine new ideas and techniques in assessment, I will argue that rubrics have the power to undermine assessment. Scoring rubrics are pivotal in operationalizing largescale and standards based performance assessments in writing. Rubrics promote reliability in performance assessments by standardizing scoring, but they also standardize writing. The standardization of a skill that is fundamentally self-expressive and individualistic obstructs its assessment. And rubrics standardize the teaching of writing, which jeopardizes the learning and understanding of writing. (p. 674).	A large number of the excerpts are based on anecdotal evidence not backed up by scientific empirical evidence. The author is an expert on assessment so she has ample experience. However, most of the claims (e.g. excerpts 11 and 15) are to rubrics in accountability approaches to testing. The problem is not the rubrics per se, but the standardization of writing assessment and the use of rubrics for such purposes. Other uses of rubrics, such as formative purposes, are not mentioned. Her critique seems to be based her

		6. Rubrics are artifacts of the psychometric tenet that good assessment begins with careful thinking about what a test taker should know and how that knowledge should be appraised. In standardized, norm referenced, multiple choice testing, agreements about what constitutes achievement are operationalized in a table of specifications and then in a test. Similarly, in largescale performance assessment, agreements about what constitutes achievement are operationalized in rubrics and in performance items of various types. The strategy is predictive in that the performance of test takers is anticipated, and it is preordinate in that what will count as satisfactory performance is determined before the test is administered. This is an orderly, linear conception of test development. But current practice of standards based assessment presents a contrastingly complicated scene with a tangle of ideas about standards and rubrics. (p. 674)	experience on rubrics that were commonplace in the 1990s in the USA. According to Mabry, there was a strong tendency for these rubrics to focus on low-levels of writing competence, rather than on rhetorical composition capabilities. Other claims are definitional in nature (e.g. excerpts 10 and 13), and it could therefore be argued that they do not need empirical evidence. However, some of the claims seem to present a biased perspective of the use of rubrics. For example, excerpt 14 presents a simplistic use of rubrics by the teachers, even though it can be assumed that a teacher would be able to identify aspects of good writing, even if these are not included in the rubric. Or that the teacher can use a simpler rubric version, because students need to work on just one criterion (e.g. organization of ideas). Finally, there is a high number of self-citations, personal communications, and conference papers.
	Limitations of criteria **Standardization and narrow the curriculum**	7. Producing a single score appears to be enough to satisfy rubric developers that a score is holistic. But when the awarding of an overall score is the result of aggregating subscores on components, the single final score is a product of analytic scoring. Although it is possible to derive a holistic score from list of discrete criteria, strategies that focus on criteria are fundamentally analytic in character because they focus on components. Because rubrics to assess writing prescribe the criteria by which papers are to be judged, claims of their holism rarely survive analysis. (p. 675) 8. Rubrics offer both explication and complication. Rubrics tend to improve interrater reliability, the likelihood that different raters will award similar scores. But the consistency is not achieved because rubrics provide a vehicle for expressing naturally occurring agreement. Rather, rubrics limit the scope of variability of scores. That is, rubrics improve interrater reliability partly by directing all scorers to judge student writing according to the same few criteria, a sameness that encourages agreement in scores. And writing rubrics typically incorporate attenuated measurement scales of three to six performance levels, which offer a scorer few choices. The fewer the choices, the fewer the possibilities for disagreement among scorers, and the fewer but more serious the measurement errors.* Agreement is further enhanced because scorers are trained to use rubrics uniformly and are monitored as they do so, their continued employment contingent on a record of awarding consistent scores. Although rubrics promote reliability, they may simultaneously undermine validity, the more important determinant of the quality of an assessment. Writing rubrics can fail to predict the actual features of a student's writing, thereby creating a mismatch between scoring criteria and actual performance. In cases in which the overall effect of a student performance is achieved by means not anticipated in the scoring criteria, criterial analysis of the quality of writing will deflect a scorer's attention away from the actual writing, and the score will not support valid inferences about the student's achievement. (p. 675)	

	9. Domain appropriateness. Rubrics present another potential threat to validity in the assessment of writing. Prespecification of scoring criteria is consistent with traditional principles of test development the longstanding practices of articulation of constructs, purposes, topics, formats, and difficulty levels before test construction and with the traditional practice of standardizing across test takers and contexts. But complying with prespecifled criteria and common standards works against creative self-expression, which is the essence of skillful writing. Because they restrict the flexibility scorers need in order to identify and commend unique strengths and skills, rubrics used for scoring large scale, standardized performance assessments in writing not only undermine validity but are fundamentally domain inappropriate, not sufficiently relevant to and representative of the domain of writing. (p. 675)	
	10. Choice of criteria. Writing rubrics typically feature four to six criteria that give priority to the mechanical, format, and organizational aspects of writing rather than to the more substantial aspects such as content, logic, compelling presentation, vivid description, figurative use of language, depth of character, or significance of theme. It may be that low level criteria dominate writing rubrics because agreement among scorers is more easily achieved. (p.676)	
	11. Similarly, as the crucial mechanism for determining scores in direct writing assessment, rubrics overwhelm the writing curriculum. The powerlessness of educators to resist the implacable influence of state testing in general and of writing rubrics in particular was apparent in observations of local administration of state mandated performance assessments in Pennsylvania, Michigan, and Indiana in 1998. (p. 677)	
	12. Teachers are thus trained to comply with test requirements and to make student writing conform to assessment rubrics. This is not training that encourages teachers to consider critically whether student achievement is helped or hurt by the tests or to judge whether the impact of state assessments on curriculum should be welcomed or opposed. (p. 677)	
Instrumentalism and "criteria compliance"	13. Often, it is in the name of equity that students are given copies of the rubrics by which their writing will be scored. Some rubrics are even printed in test booklets. On the face of it, it does seem reasonable and fair to provide students with information about the expectations that they will be asked to meet. It is less obvious that doing so permits rubrics to set the boundaries of creative expression. Rubrics are designed to function as scoring guidelines, but they also serve as arbiters of quality and agents of control. Moreover, the control is not limited to assessment episodes but influences curriculum choices, restricts pedagogical repertoires, and restrains student expression and understanding. (p. 678)	
Limitations of criteria	14. The move toward direct writing assessment is a move in the direction of authenticity and validity. Clearly, standardized multiple-choice testing presents a validity problem for assessing writing. Multiple-choice items about different aspects of writing incorrectly imply that good writing is the sum of such components as spelling, vocabulary, grammar, word choice, and sentence structure and that the ability to answer multiple-choice items on these topics is a measure of the ability to write well. But standardizing scoring with a writing rubric also presents a validity problem. Rubrics incorrectly imply that good writing is the sum of the criteria on the rubric, that the criteria on the rubric are sufficient for good	

		writing, and that writing that does not conform to the criteria on the rubric is not good. (p. 678)	
	Standardization and narrow the curriculum	15. Standardization is essential to large scale testing. Standardization of direct writing assessments promotes reliability in scoring and facilitates the comparison of students, schools, and districts. Thus standardization caters to the insatiable public appetite for rankings. And there's the rub. Rubrics standardize scoring, and so they standardize writing. But standardized writing, by definition, is not good writing because good writing features individual expression, which is not standardized. The standardization of any skill that is fundamentally individual obstructs its assessment. (p. 678)	
Norton, 2004	**Instrumentalism and "criteria compliance"**	16.Helping students to concentrate on assessment criteria paradoxically means that they may take a strategic approach and end up focusing on the superficial aspects of their assessment tasks, rather than engaging in meaningful learning activity. (p. 687)	Empirical research. Criteria are used as part of a case study on the development of PBL methodology in a counselling psychology module. Empirical data are student evaluations.
		17. Unfortunately, adding more detail about the assessment criteria in this way had the unintended effect of apparently raising students' anxieties and leading them to focus on sometimes quite trivial issues. /…/ This experience has challenged the author's assumptions that a focus on making assessment criteria explicit will enable students to take a deep approach. (p. 693)	
O'Donovan, Price & Rust, 2004	**Simple implementations don't work**	18. a) /…/ despite our best efforts, on their own, the explicit assessment criteria and grade descriptors failed to transfer meaningful knowledge on assessment standards and criteria to students." (p. 327) b) Meaningful knowledge of assessment and standards is best communicated and understood through the use of a combination of both explicit and tacit transfer processes. (p. 333)	Empirical research. "/…/ own research undertaken with large classes of students (300+) /…/ Students in their first term at university are invited to a 90-minute 'marking workshop', prior to which they must mark and give feedback on two unmarked exemplar assignments (previously tutor marked as 'A' grade and borderline). The assignments are similar in nature and format with the same marking criteria as that of their next written assessment, but are different in task and topic. /…/ Our findings (replicated for 3 years) show students who undertake this optional marking workshop demonstrate a significant improvement in performance compared with those who do not, even though base line comparison of the performance of the two groups, undertaken prior to the intervention, shows no significant difference in performance" (p. 332) The term rubric is only used once in the paper. The term " criterion assessment grid" is used instead (p. 327).
Andrade, 2005	**Simple implementations don't work**	19. Rubrics are not entirely self-explanatory. Students need help in understanding rubrics and their use. When I once handed out a rubric that we had cocreated in class and assumed that students knew what to do with it because we had cocreated it, I was in for a surprise. The more motivated students anguished over what to do with it and the less motivated filed it in their notebooks and promptly forgot about it. (p. 29)	A mostly theoretical and practical paper coming from the author's experience as researcher and teacher. However, this is clearly stated in the first paragraph and, therefore the reader is advised.
		20. Similarly, rubrics are not a replacement for good instruction. Even a fabulous rubric does not change the fact that students need models, feedback, and opportunities to ask questions, think, revise, and so on. (p. 29)	

	Standardization and narrowing the curriculum	21. Issues of validity, reliability, and fairness apply to rubrics, too. We need to worry more about the quality of the rubrics that we use. I have seen some very idiosyncratic rubrics in my day, and this is where it gets ugly. “Alternative” or “authentic” assessments are not exempt from the demands of validity, reliability, and equity (Moskal and Leydens 2000)...At a minimum, an instructional rubric must be aligned with reasonable and respectable standards and with the curriculum being taught in order to be valid. It must pass a test of reliability by resulting in similar ratings when used by different people. (p. 29)	
Andrade, 2006	**Instrumentalism and "criteria compliance"**	22. Kohn worries that rubrics focus attention on only the most quantifiable and least important qualities of assignments. He points out that rubrics that emphasize spelling and organization result in "vacuous writing." (p. 9)	It is a reply to Kohn (2006) –see below-. It does not contain empirical evidence to support the critique to rubrics as it is actually more a “defense of rubric” publication.
		23. Kohn also complains that rubrics can become assignment maps that students mindlessly follow. He quotes a sixth grader who says, "The whole time I'm writing, I'm not thinking about what I'm saying or how I'm saying it. I'm worried about what grade the teacher will give me . . . " (14). (p. 9)	
Kohn, 2006	**Miscelanea**	24. The ultimate goal of authentic assessment must be the elimination of grades. But rubrics actually help to legitimate grades by offering a new way to derive them. They do nothing to address the terrible reality of students who have been led to focus on getting A’s rather than on making sense of ideas. (p. 12)	This publication presents anecdotal evidence, some transfer of empirical results from other fields to make an argument (e.g. Dweck), and cites the work by Mabry (1999) and Wilson (2006) analyzed above which do not follow a systematic scientific inquiry. It can be concluded Kohn (2006) does not contain scientific empirical evidence to support the claims only anecdotal one.
	Standardization and narrow the curriculum	25. Finally, there’s the matter of that promise to make assessment “quick and efficient.” I’ve graded enough student papers to understand the appeal here, but the best teachers would react to that selling point with skepticism, if not disdain. They’d immediately ask what we had to sacrifice in order to spit out a series of tidy judgments about the quality of student learning. (p. 12)	
	Limitations of criteria	26. Now some observers criticize rubrics because they can never deliver the promised precision; judgments ultimately turn on adjectives that are murky and end up being left to the teacher’s discretion. (p. 13)	
	Standardization and narrow the curriculum	27. But I worry more about the success of rubrics than their failure. Just as it’s possible to raise standardized test scores as long as you’re willing to gut the curriculum and turn the school into a test-preparation factory, so it’s possible to get a bunch of people to agree on what rating to give an assignment as long as they’re willing to accept and apply someone else’s narrow criteria for what merits that rating. Once we check our judgment at the door, we can all learn to give a 4 to exactly the same things. (p. 13)	
		28. To this point, my objections assume only that teachers rely on rubrics to standardize the way they think about student assignments. Despite my misgivings, I can imagine a scenario where teachers benefit from consulting a rubric briefly in the early stages of designing a curriculum unit in order to think about various criteria by which to assess what students end up doing. As long as the rubric is only one of several sources, as long as it doesn’t drive the instruction, it could conceivably play a constructive role. (p. 13)	

		29. Just as standardizing assessment for teachers may compromise the quality of teaching, so standardizing assessment for learners may compromise the learning. Mindy Nathan, a Michigan teacher and former school board member told me that she began “resisting the rubric temptation” the day “one particularly uninterested student raised his hand and asked if I was going to give the class a rubric for this assignment.” She realized that her students, presumably grown accustomed to rubrics in other classrooms, now seemed “unable to function unless every required item is spelled out for them in a grid and assigned a point value. Worse than that,” she added, “they do not have confidence in their thinking or writing skills and seem unwilling to really take risks.” (p. 13)	
	Instrumentalism and "criteria compliance"	30. The fatal flaw in this logic is revealed by a line of research in educational psychology showing that students whose attention is relentlessly focused on how well they’re doing often become less engaged with what they’re doing. There’s a big difference between thinking about the content of a story you’re reading (for example, trying to puzzle out why a character made a certain decision), and thinking about your own proficiency at reading. (p. 13)	
		31. To that extent, more detailed and frequent evaluations of a student’s accomplishments may be downright counterproductive. As one sixth grader put it, “The whole time I’m writing, I’m not thinking about what I’m saying or how I’m saying it. I’m worried about what grade the teacher will give me, even if she’s handed out a rubric. I’m more focused on being correct than on being honest in my writing.”[8] In many cases, the word even in that second sentence might be replaced with especially. But, in this respect at least, rubrics aren’t uniquely destructive. Any form of assessment that encourages students to keep asking, “How am I doing?” is likely to change how they look at themselves and at what they’re learning, usually for the worse. (p. 14)	
	Standardization and narrow the curriculum	32. What all this means is that improving the design of rubrics, or inventing our own, won’t solve the problem because the problem is inherent to the very idea of rubrics and the goals they serve. This is a theme sounded by Maja Wilson in her extraordinary new book, Rethinking Rubrics in Writing Assessment.[9] In boiling “a messy process down to 4-6 rows of nice, neat, organized little boxes,” she argues, assessment is “stripped of the complexity that breathes life into good writing.” High scores on a list of criteria for excellence in essay writing do not mean that the essay is any good because quality is more than the sum of its rubricized parts. To think about quality, Wilson argues, “we need to look to the piece of writing itself to suggest its own evaluative criteria” – a truly radical and provocative suggestion. (p. 14)	

Wilson, 2006	**Instrumentalism and "criteria compliance" & Standardization and narrow the curriculum**	33. a) She made allegations, mostly coming from personal experiences of herself or colleagues, about how rubrics decrease creativity and quality of writing "What were many rubrics, really, but ultimate graphic organizers-clearly defined performance levels organized into clearly defined boxes in a chart". (p. 10) b) "When our purpose in reading student work is to defend a grade, we do not apply any of our natural responses to text. Encouraged by the performance levels on the rubric to rank students against an external standard, our readings of student work are based firmly in a deficit model" (p. 30) "Essentially, any rubric with numbers attached to various performance levels is a mathematical formula or model that determines how different isolated factors of writing work together to create the effectiveness of a piece of writing; we plug in the numbers, and it spits out the grade." (p. 32) "The rubric's grading mechanism depends on the laws of determinism, of simple linear cause and effect. Rubrics claim that if we just get the factors and weightings right, teachers can plug in the numbers and the rubric will reliably predict good or bad writing". (p. 32) "Rubrics attempt to fix our confusion about responding to student papers in several ways. First, rubrics prevent teachers from turning less is more into nothing is more. Even if a teacher takes no time to write comments but uses the 6 + 1 Trait(c) rubric and gives a score of 3 for sentence fluency, the student will know that " Although sentences may not seem artfully crafted or musical, they get the job done in a routine fashion". (p. 35) "While rubrics sketched a rough terrain of the factors of writing that was helpful for early composition theorist and teachers, it is not a full or rich enough map for our use today". (p. 36) "If rubrics, developed in the mid-twentieth century, are based on a limited notion of good writing, then we hold students to an outdated notion of good writing when we use them today. In a 1984 study by Sarah Freedman, students and professional writers were given the same writing prompt (cited in Huot 2002a). When their work was judged using a rubric, the professional writers scored lower than the students; the rubric did not honor the sophistication and variety of approaches that the professional writers brought to bear on the prompt". (p. 37) "Rubrics direct me to read in a way that drains the meaning and joy from teaching writing". (p. 39) "While our use of rubrics most obviously damages the student writers who refuse to fit into our molds, "good students" suffer under the reign of rubrics as well". (p. 39) "Rubrics encourage us to read and our students to write on autopilot" (p. 39) "The reductive categories of rubrics don't honor the complexity of what we see in writing and what our students try to accomplish". (p. 41)	The majority of claims come from the writer's personal experience or other articles based on personal experiences. Regarding the particular excerpts included in the column on the left, only one reference to empirical research is made (i.e. in relationship to Freedman (1984) and Huot (2002)). When exploring these references, the original article by Freedman could not be accessed. Instead, the explanation given in Huot (2002) about the Freedman reference was explored. However, the claims made by Wilson could not be substantiated.

Torrance, 2007	**Instrumentalism and criteria compliance**	34. The clearer the task of how to achieve a grade or award becomes, and the more detailed the assistance given by tutors, supervisors and assessors, the more likely candidates are to succeed. But transparency of objectives coupled with extensive use of coaching and practice to help learners meet them is in danger of removing the challenge of learning and reducing the quality and validity of outcomes achieved. This might be characterized as a move from assessment of learning, through the currently popular idea of assessment for learning, to assessment as learning, where assessment procedures and practices come completely to dominate the learning experience, and 'criteria compliance' comes to replace 'learning'. (p. 282)	Empirical research. A series of parallel case studies of assessment "in action" across a wide variety of LSS settings and by a questionnaire distributed to a larger sample of learners derived from the case study settings. Case studies were conducted of assessment across of the post-compulsory "Learning and Skills Sector" (e.g. further education colleges, workplaces and adult learning environments). Findings show: (1) Achievement is routinely defined in fairly narrow and instrumental terms; (2) Failure is defined as non-completion or not securing expected and necessary grades; (3) Students can re-sit modular tests to improve grades or retrieve fails; (4) There is a significant, even overwhelming, culture of support for learners/candidates at every level, for instance by tutors breaking down and interpreting assessment criteria; (5) Students can draft and re-draft assignments, receiving feedback on strengths and weaknesses and what needs to be done to improve the grade; (6) Support can be observed in the way "leading questions" are asked of candidates to help them through observations of workshop practice and in compiling portfolio evidence.
Wilson, 2007	**Standardization and narrow the curriculum**	35. The idea that we can standardize our responses to students' papers deserves serious examination, because language itself resists all but the most basic attempts to standardization. (p. 63) 36. I suggest that we make ourselves transparent as we read - that we pay attention to what goes on in our minds and try to put our reactions and questions and wonderings and musings and connections and images into words (p. 63)	Anecdotal evidence: The claims are based on the author's own experience as a teacher, with no scientific empirical evidence for the claims
Chapman & Inman, 2009	**Standardization and narrow the curriculum**	37. Leaving aside the inherent possibility that the child might have misrepresented her teacher's intent slightly, we were struck by a child's veracity about the restrictions a rubric-oriented teaching force places on our learners. Such restrictions may be real: the students must adhere strictly to prescribed criteria with no deviations, per the teacher's instructions, or student culture may impose restrictions (i.e., "Those criteria represent all I have to do in order to have a 'perfect' assignment"). (p. 198)	Anecdotal evidence (see excerpt below this paragraph). The authors assert to have empirical data for their claims, such as quotes from students, discussions with peers, and reviews of sample rubrics. However, there is no methodological information stating, for instance, how the students or rubrics were sampled or analyzed. "When we discussed the matter with fellow college of education faculty members, we were told - here may be a key concept - that "well-constructed" rubrics do allow for the creativity and flexibility we believe imperative in meaningful learning. Reviewing sample rubrics used in thirty undergraduate courses, however, we found little flexibility and even less encouragement of individual initiative: almost every rubric, along with samples of student work, pointed to

	Instrumentalism and "criteria compliance"	38. Hand-in-hand with this first example is the question of whether teachers set the bar low with such scoring rubrics so that all students reach the bar. Given the wide range of achievement levels that exists in any given class, some students will grasp concepts and some will merely grasp at concepts. A teacher must purposely set the passing bar low if the great majority of the class population is to demonstrate competency. Should teachers set a minimally acceptable product as the middle rung on their scoring rubric because that is all students need demonstrate to meet national, state, or local "standards," and work both up and down from there? If teachers use such scoring rubrics to assign grades on products, how receptive will they be to a majority earning less than a high score? Likely not at all. If, instead, a minimally acceptable product is the top rung, then how close to the standards does a mere "C" come? And more alarmingly, what incentive is there for any student to go beyond what is required to simply clear the bar or achieve whatever competence level they decide on, based on rubric minimums? (p. 199)	a bland sameness. Standardization and uniformity seemed honored and, by extension, desired. If that is what we are modeling for our future teachers, what can we realistically expect them to put into practice? (p. 200)".
Sadler, 2009	**Limitations of criteria**	39. /…/ the portrayal to students is that analytic grading first involves a series of judgments made at the level of individual criteria. After these are completed, they are processed into a grade using a well-defined algorithm in an essentially linear sequence of steps. (p. 164)	Anecdotal evidence as shown by this excerpt: "In this section, six types of significant non-standard events are identified and described under the heading Observations. They manifest themselves across a wide range of university teachers, in a wide range of disciplines and fields, and for many types of assessment tasks and student works. Each observation has been investigated in conversations with academics, then elaborated and refined." (p. 164)
		40. But as many teachers testify, the actual process is much more convoluted. Markers generally do not scrutinise an entire work multiple times, once for each criterion. As mentioned earlier, they run simultaneously with dual agendas, one being to develop a perspective on the work's overall quality, and the other to take note of particular characteristics or deficiencies that are worthy of special focus or attention. These particular 'noticings' are useful both in shaping the emerging perspective and in generating raw material for reporting. As most academics can attest, working systematically through the criteria one at a time would be highly labour intensive. Apart from the labour aspect, a significant reason for not operating in this way is that assessors are initially more interested in how the work comes together as a whole than in performance on individual criteria. (p. 164-165)	
		41. During the process of grading certain works, experienced assessors often become aware of discrepancies between their global and their analytic appraisals. A work judged as 'brilliant' overall may not rate as outstanding on each criterion. This would be necessary, logically and arithmetically, for the work to be assigned the top grade. Conversely, another work that comes out well on each criterion may be judged as only mediocre overall. (p. 165)	
		42. When assessors identify works for which the analytic and global judgments differ, they may or may not be able to account for the discrepancy. Many teachers can identify with situations in which a work exhibits an indefinable 'quality,' inherent in its wholeness, which simply cannot be passed over as irrelevant or inconsequential. /…/ in other cases the assessor knows the reason: it lies in a positively identified criterion which was not included in the preset list distributed to students but turns out to be crucial to the judgment. (p. 166)	
		43. A necessary condition for any formal system that uses discrete criteria as logical entities in their own right is that the criteria are conceptually distinct from one another. Each criterion is assumed to have an established interpretation that, at least in theory, represents	

		a property that is different from those signified by the other criteria, taken singly or together. (p. 166-167)	
		44. When the criteria are treated as separate variables, each criterion stands on its own. No account is taken of situations in which the co-occurrence of particular levels on two or more criteria contributes more, or less, to the overall quality determination than is reflected in the separate levels, either singly or when combined. Co-occurrence on two criteria is known as a two-way interaction. The typical analytic grading scheme does not allow for interactions at all, yet they can be pivotal to certain grading decisions… (p. 172)	
		45. Different university faculty regularly use different sets of criteria, even for student works in the same genre. /…/ In many contexts it is not practical to nominate all the criteria that could conceivably be used. Choices therefore have to be made. Selecting particular criteria is a decision to exclude others. (p. 167)	
	Standardization and narrow the curriculum	46. In many assessment situations, the set of specified criteria is actually a sample from a larger pool or 'population' of criteria. Such selections are not random samples, of course. They are usually made carefully, but the characteristics of samples do not match precisely those of their corresponding populations. In the current context, it follows that any selection of criteria cannot possess the same rich appraisal potential as does the population. /…/ bias is necessarily introduced by this narrowing of scope /…/ with validity lowered as a result. (p. 169)	
	Limitations of criteria	47. Different assessors often agree on the overall grade for a particular work, but differ substantially about levels of performance on the separate criteria. (p. 167)	
	Context-dependence	48. Some properties of likely interest to academic assessors are concrete and unambiguous, including length (of a written work), or conformity with an external convention (such as referencing style). /…/ However, a large proportion of assessment criteria are not at all like these. They refer to concepts which are denoted primarily by words and their meanings, not to measurable properties. In ordinary discourse, words need to be interpreted in context. (p. 169)	
		49. A consequence of the lack of unique meanings is that, within the same context, criteria may be interpreted differently by different teachers. They can also be interpreted differently by the same teacher in different assessment contexts. (p. 169)	
		50. See also: The meanings for the principal elements, qualities and amounts alike, are inherently context dependent. They are not and cannot be standardised. Since contexts differ, a stand-alone codification cannot be interpreted in a unique way by different people in different contexts at different times. (Sadler, 2014, p. 275)	
	Limitations of criteria	51. Some evaluative characteristics may actually be impossible to articulate, even in principle. 'Criteria' that can only be 'tacitly known' pose a significant communicative challenge. (p. 170)	
		52. By claiming to be objective and open, they project strongly on both ethical and practical fronts. /…/ They promise more than they can deliver. (p. 172)	

Rezaei & Lovorn, 2010	**Simple implementations don't work**	53. It should be noted; however, that simple implementation of rubrics may not guarantee effective assessment (Breland, 1983; Ross-Fisher, 2005; Tomkins, 2003). (p. 19)	Breland (1983) refers to rubrics twice, but not in a critical way. Tomkins (2003) does not use the term rubric, but mentions "marking schemes". She describes her experience with the scheme, raising doubts as to the validity of the scores (anecdotal evidence). Ross-Fisher (2005) is a publication in favor of using rubrics. In sum, none of the references offers empirical evidence for the claim not even they seem to offer theoretical support as they do not cover the topic.
		54. Teachers, schools, and school systems have adopted rubrics for more accurate assessment in every discipline. Recently, however, some educators have challenged the collective assumption that simply implementing rubrics increases inter-rater reliability and validity, and the overall accuracy and quality of assessment (Kohn, 2006; Wilson, 2006). (p. 20)	None of the references present scientific empirical evidence.
	Standardization and narrow the curriculum	55. In steadily increasing numbers, educators are coming to realize that no rubric can be completely effective in evaluation of students' individual writing idiosyncrasies or their unique understanding of the concepts. Some have even found that rubrics prematurely narrow and cement their visions of good writing (Wilson, 2007). (p. 20)	That reference does not present scientific empirical evidence.
Bloxham, Boyd & Orr, 2011	**Limitations of criteria**	56. The study found that assessors made holistic rather than analytical judgements. A high proportion of the tutors did not make use of written criteria in their marking and, where they were used, it was largely a post hoc process in refining, checking or justifying a holistic decision. Norm referencing was also found to be an important part of the grading process despite published criteria. (p. 655)	Empirical research. Twelve lecturers from two universities were asked to "think aloud" as they graded two written assignments. The study found that assessors made holistic rather than analytical judgements. A high proportion of the tutors did not make use of written criteria in their marking and, where they were used, it was largely a post hoc process in refining, checking or justifying a holistic decision. Norm referencing was also found to be an important part of the grading process despite published criteria.
		57. There is a disjunction between stated policies and actual practices in higher education marking, particularly in relation to analytical, criterion referenced grading. (p. 667)	
Lovorn & Rezaei (2011)	**Simple implementations don't work**	58. Research also revealed, however, that although teachers and administrators may perceive rubrics as inherently reliable (Jonsson & Svingby, 2007; Silvestri & Oescher, 2006), these instruments do not guarantee effective assessment (Ross-Fisher, 2005; Tomkins, 2003). Mabry's study (1999) even suggested that rubrics may sacrifice validity to increase reliability. (p. 2)	Ross-Fisher (2005) does not provide any support for the claims made by the authors. Tomkins (2003) provides only anecdotal evidence. Mabry (1999) also relies heavily on anecdotal evidence.
		59. More recently, increasing numbers of teachers, administrators, and researchers have challenged collective assumptions that simple use of rubrics leads to increases in inter-rater reliability, evaluation accuracy, and/or quality of assessment (Chapman & Inman, 2009; Dawson, 2009; Kohn, 2006; Reddy & Andrade, 2010; Stellmack, Konheim-Kalkstein, Manor, Massey & Schmitz, 2009). (p. 2)	A mix of different types of papers are used to support a general idea (i.e. that rubric reliability is been contested). For example, Stellmack et al. (2009) contains empirical evidence about the claim. Chapman and Inman (2009) is a purely informational text discussing similar ideas. Kohn (2006) has been analyzed above and also address similar aspects. In sum, those references back up the statement that aspects of reliability are contested.
	Standardization and narrow the curriculum	60. Research indicates that more educators hold the opinion that rubrics, in and of themselves, offer no guarantee of effective evaluation, particularly in terms of students' individual writing idiosyncrasies or their unique understanding of concepts (Cooper & Gargan, 2009; Lumley & McNamara, 1995; Malouff, 2008); and may even narrow and bias	Cooper and Gargan (2009) is informational and anecdotal. Lumley and McNamara (1995) do not use rubrics. In their study, rating scales were used, which is a different tool than rubrics (Brookhart, 2014).

		raters' visions of good writing (Read, Francis & Robson, 2005; Schafer, Gagné & Lissitz, 2005; Tomkins, 2003; Wilson, 2007). (p. 2)	Malouff (2008) it is an anecdotal and informational paper. However, it states the opposite of what Lovorn and Rezaei are using the reference for. Malouff intended to use rubrics to reduce his bias when scoring. Read et al. (2005) is an empirical paper, but rubrics are not mentioned. The only similar term used is "the marking scheme at their own institution". Furthermore, the only statement in Read et al. containing this term, and including content related to the claim made by Lovorn and Rezaei, is the following: "Practices such as external moderation and the introduction of standardized marking schemes are likely to play a sizeable role in reducing the amount of variation in assessor judgements (although a number of participants in our study mentioned that when assessing essays they first make an assessment without reference to their departmental marking scheme, and then 'fit' their judgement to the scheme's categories—see Francis et al., forthcoming)." (p. 257). This statement does not support the claim made by Lovorn and Rezaei, but rather that assessors tend to ignore the marking scheme and rely on their initial holistic judgment. Schafer et al. (2005) explore variables that could influence the scoring in a large-scale test. No support could be found for the claim made by Lovorn and Rezaei are claiming. Both Tomkins (2003) and Wilson (2007) provide only anecdotal evidence.
	Simple implementations don't work	61. Teachers' misuses, biases, and inconsistencies related to rubrics may be due to inadequate training. Turley & Gallagher (2008) suggested that teachers untrained in rubric purpose, design, and implementation often use the instruments improperly, rely on them too much, or see no value in them at all. Wilson (2007) concluded that many poorly trained teachers use rubrics in ways that compartmentalize and bias their evaluations of students' reading and writing skills. Additionally, Rezaei & Lovorn (2010) found that poorly trained or untrained teachers who use of rubrics to assess students' writing submissions are significantly less consistent in their evaluations than those who receive adequate or good training on rubric construction and use. (p. 2)	Turley and Gallagher (2008) and Wilson (2007) are based on anecdotal evidence. Furthermore, Turley and Gallagher (2008) is a publication in favor of using rubrics. Rezaei and Lovorn (2010) provides empirical evidence, but has serious validity threats in their interpretation of the results.

Shipman, Roa, Hooten & Wang, 2012	**Limitations of criteria**	62. Researchers have found limitations to utilizing rubrics when depicting performance expectations. Consequently, this creates issues with educator bias and unreliable scores (Gantt, 2010; Knight et al., 2010; Kohn, 2006). (Authors note: the following belongs to the same paragraph in the original Shipman et al. However, as the reader can note, the critique shifts from rubrics to other assessment aspects with no further reference to rubrics. This can be a case of incorrect formulation as the reader can be left with the sensation the authors are still referring to rubrics). The issues of educator bias regarding what constitutes an acceptable performance can undermine the performance standard. Nash and Lewandowski (2010) found faculty who know and understand the students' emotional lives result in a productive learning environment. On the contrary, educators who know their students may intertwine personalities and preferences generating biased and subjective judgments especially when criteria are not explicit (Andrade, 2005; Murphy, 2004; Neumann and Forsyth, 2008). In addition, Jae and Cowling (2009) found that if the grader knows the student being assessed, then bias in grading will be pervasive. (p. 268)	Gantt (2010) provides empirical evidence, but has severe flaws in methodology and report. For example, this claim is made: "We found that the verbs from Bloom et al. (1956) used in the Clark rubric may have any number of definitions as considered by faculty. In some cases, faculty rated students between categories as opposed to placing student performance clearly into one category.", which is the followed by this: "We did not endeavor to establish interrater reliability during the initial applications of the rubric; this is work we hope to undertake as we continue to make decisions about which tools work best for student simulation evaluation." A as consequence, this reference does not support the claim by Shipman et al. (2012). The article by Knight et al. (2010) could not be accessed in full, only an extended abstract was available. This paper does not, however, present an empirical investigation, but demonstrates how a particular method (”the six sigma method”) can be used to establish the reliability of a rubric Kohn (2006) provides mostly anecdotal evidence. Nash & Lewandowski (2010) provide empirical evidence, but in relation to “pre-lesson reflection questions”, not rubrics. Andrade (2005) is a theoretical and practical paper providing anecdotal evidence. Murphy (2004) did not use a rubric, but a 5-point Likert scale. Neumann and Forsyth (2008) did not use any rubrics. Jae & Cowling (2009) do not focus on rubrics, but on identifying students by using bar codes in order to assure student anonymity and reduce bias.

		63. However, if the rubric is poorly constructed it can complicate issues in the areas of expectation and grading. Creating a rigid rubric without flexibility allows for interpretation by the faculty which can limit the generalizability of the tool (Bresciani et al., 2009)...Isaacson and Stacy (2009) found issues when the rubric is used for clinical evaluations. The shortcoming of many clinical evaluations is the lack of describing the performance levels with respect to attributes in the affective and psychomotor domains of learning. Additionally, there are criticism rubrics which can never deliver the promised precision and judgments; which end up being left to the educator's subjective discretion (Hitt and Helms, 2009; Knight et al., 2010; Kohn, 2006). For example, Gantt (2010) found that faculty would have a mental picture of a passing student and would adjust the rubric scoring accordingly making the grading process fair. (p. 268)	Bresciani et al. (2009) provide empirical evidence suggesting that rubrics, even without training, enhance inter-rater reliability. It might be that Shipman et al., were referring to this excerpt from Bresciani: "However, some researchers warn that including rigid definitions in rubrics might limit their generalizability (Colton et al., 1997)." This probably comes from the original Colton et al.: "Also, it may be possible to increase rater consistency by more rigidly defining scoring rubrics, but again, this might limit the generalizability." (page 5), which is done in a section about definitions of reliability and does not appear to refer to any empirical evidence. In sum, the claim about Bresciani et al. by Shipman et al. is not based on scientific empirical evidence. Isaacson & Stacy (2009) is a paper in favor of rubrics (with no scientific empirical evidence). They conclude: "They [Rubrics] can be all encompassing or encourage lengthy comments. Although tedious to develop, the process itself is very educational for those involved and the end product is an overall improved assignment for faculty and students alike. Faculty become vested in all aspects of their assignments when they walk through the process of rubric development and students know what to expect and what is expected of them. Rubrics are a valid answer to the concerns of faculty and students related to clinical evaluation." Hitt and Helms (2009) is a paper in favor of rubrics, presenting rubrics as a solution for teacher bias (although the paper does not present any empirical evidence for the claims). The authors only have one paragraph about shortcomings, which seems to be the one picked up by Shipman et al.: "However, we also concede that rubrics cannot eradicate all sources of biases in teaching or schooling. The rules and regulations for school systems and individual schools can create favored or advantaged students. For example, some students may be identified as ’gifted‘ or ’advanced,’ and other students can be identified as ’at-risk.’ Unfortunately, this labeling process can be susceptible to personal biases." However, the rest of the paper is in opposition to the claims made by Shipman et al.: Hitt & Helms assert that rubrics are a great help to avoid teachers' assessment bias. As mentioned earlier: Knight et al. (2010) present a demonstration of a particular methodology, while Kohn (2006) provides a theoretical critique with anecdotal evidence. Gantt (2010) mentioned: "Because of faculty inexperience in using the clark rubric, no definitive passing score or level was determined for use with the aDn students. The community college faculty were committed to having each student complete the scenarios, but

			wanted to avoid the anxiety associated with a numeric grade. However, in working with the rubric with this group of students, faculty began to experiment with methods of determining a numeric score for use with the rubric." /…/ "faculty or staff who will be scoring the scenarios must meet ahead of time to discuss how students can demonstrate adequate performance; a determination of what constitutes a passing score must also be made. Faculty members with disparate teaching styles can undermine reliability, as in the case when the evaluative scenario turns into the teaching scenario where students' actions are interrupted." /…/ "We did find that faculty may tend to develop a picture of the passing student within their minds, and then adjust the rubric scoring accordingly." /…/ "Knowing that even a graduating student will likely not receive a score beyond the second column of the rubric or the advanced beginner level, the faculty member may have difficulty determining a satisfactory or passing score based on a total of 100 percent." As mentioned earlier, there are some methodological flaws in Gantt because, as can be read in the first excerpt from this paragraph, it was decided not to associate the performance with a numeric grade, but then it seems the faculty did use the rubric to score and pass the students. Therefore, it is not clear what part was the "real" learning situation and what the "empirical" situation. Furthermore, the reporting of results is rudimentary and does not comply with APA guidelines. Consequently, it is questionable how much weight should be attributed to the empirical findings from this study.

		64. For example, in 1985 Pace contended that the difficulty with student assessments is professors have different criteria for evaluating students. As more instructors use the rubric tool, a decrease in inter-rater reliability may be evident. One reason is that objectives may be unclear with structured categories. Therefore, the result is left to the instructor's discretion (Chapman and Inman, 2009; Kohn; Moskal and Leydens, 2000). Moreover, rubrics are subjective "when it is used to convert qualitative terms, each critical and independent, into a set of scores that can be summarized, averaged, and transferred into a grade" (Cooper and Gargan, 2009, para 12). (p. 268)	The part of the paragraph making reference to Pace (1985) is not about rubrics. The publications by Chapman & Inman (2009), Kohn (2006 [since no date is provided, the 2006 publication is assumed, since it is present in the reference list]), and Moskal & Leyens (2000) are all theoretical contributions, providing anecdotal evidence. Cooper & Gargan (2009) is a short introductory paper about rubrics, providing no empirical evidence for any of the positive and negative claims. Furthermore, although the meaning is the same, Shipman et al. seem to have changed the original quote: "This is especially true when rubrics are used to convert lists of qualitative terms, each critical and independent, into a set of scores that can be summed, averaged, and transformed into a grade."
	Standardization and narrow the curriculum	65. In addition, critics contend that rubrics promote conformity and standardization that are incongruent with the concept of student centered learning (Kohn, 2006). The issue with the rubric as an assessment tool can stem from a narrow interpretation from when it is used for grading rather than support of understanding (Goodrich-Andrade, 2006). Even though the rubric is a formalized evaluation method, there are issues such as fairness, reliability, and validity. The rubric should pass the test of reliability when used by different people (Andrade, 2005). (p. 268)	In relation to this excerpt, the authors cite Kohn, which it is a theoretical paper, providing anecdotal evidence, and then cite Andrade's work as negative towards rubrics, even though she is one of the major proponents of rubrics. Additionally, they do not indicate that Andrade (2006) [and not Goodrich-Andrade (2006) as they referenced it], is written as a response to Kohn (2006). None of two articles provide any empirical evidence to support the claim. Andrade (2005) does not present empirical evidence, and the original excerpt is: "Issues of validity, reliability, and fairness apply to rubrics, too....At a minimum, an instructional rubric must be aligned with reasonable and respectable standards and with the curriculum being taught in order to be valid. It must pass a test of reliability by resulting in similar ratings when used by different people." (p 29-30).
Torrance, 2012	**Instrumentalism and criteria compliance**	66. /…/ we have moved towards transparency of objectives and assessment criteria, coupled with clear feedback being provided in relation to these criteria, but such a combination of transparency and feedback may not really be considered sufficient to the purpose of higher education. What we have here is not so much formative assessment, but *conformative* assessment. (p. 332)	Theoretical paper where both theory and empirical research are discussed. Most of the discussion is based on a special issue of "Assessment and Evaluation in Higher Education", focusing on formative feedback in higher education ("Approaches to Assessment that Enhance Learning": AEHE, 35, 2010). The paper synthesizes previous research, but not in a systematic manner (i.e., it is not a systematic review).
Bell, Mladenovic & Price, 2013	**Instrumentalism and "criteria compliance"**	67. Students in our study either used the exemplars, grade descriptors and marking criteria as tools for reflection and learning, or to focus on the mechanics of the assessment task. (p. 781)	Empirical research. This paper provides insight into the perceptions of first-year students of the usefulness of grade descriptors, marking criteria and annotated exemplars. Of the 119 students who provided their reflections on the resources, 87% found the resources to be useful. Students' responses about the usefulness of the resources revealed two main standpoints: those (1) seeking precise guidance and (2) happy with 'an idea' of standards.

Humphry & Heldsinger, 2014	**Standardization and narrowing the curriculum**	68. Messick (1994) raised the question of whether rubrics validly meet the purposes of their usage, asking the following: “By what evidence can we be assured that the scoring criteria and rubrics used in holistic, primary trait, or analytic scoring of products or performances capture the fully functioning complex skill?” (p. 20). Nearly two decades later, even though the use of rubrics is now widespread across the globe, surprisingly little empirical research has been devoted to answering this question (Reddy & Andrade, 2010; Rezaei & Lovorn, 2010). (p. 253)	This question still needs to be addressed by scientific empirical research.
		69. On the basis of the empirical evidence, we will argue that the widely used matrix design of rubrics can create a threat to valid performance assessment. The threat arises because there is typically no underlying developmental or learning theory that justifies having the same number of qualitative gradations across criteria. The focus here is on construct validity, as this term is defined later. We aim to show that rethinking the structural design features of rubrics may avoid this specific threat to validity by allowing rubrics to more faithfully capture qualitative gradations of performance independently for each criterion. Our intention is not to claim that resolving the threat to validity addresses other validity-related issues (e.g., whether the task and criteria are appropriate to assess a trait). However, we propose that resolving the validity threat opens the way for more productive research into a number of questions, such as to ascertain which and how many criteria should be used, whether the operational independence of criteria can be established, and the optimal number of qualitative gradations for each separate criterion. Resolving the threat to validity might also open the way to more productive research into whether raters make more valid assessments using rubrics than holistic judgments. (p. 253)	Empirical research. The design of rubrics with "forced" same number of performance levels can create validity threats.
		70. We hypothesize that the structural alignment of categories can produce an apparent halo effect for two reasons. The first reason is that structural alignment, where criteria have equal numbers of categories, may result in more or less categories for any given criterion than is optimal given the number of qualitative distinctions that raters can make. If there are too many categories, judges may have little choice but to make spurious distinctions either by defaulting to a pattern of common scoring (akin to a response set) or through recourse to a global judgment. (The issue is not the use of a global judgment per se but rather that repetition of a global judgment is contrary to the aim of analytic scoring.) If, on the other hand, there are too few categories, judges are prevented from making distinctions they are capable of making. In this case, again, ratings do not reflect variation in the quality of performances that raters can discern. The second reason is that structural alignment can create a degree of unintended conceptual overlap and redundancy in the descriptions of gradations for some pairs of criteria as described by Sadler (2009, p. 169). (p. 256)	Empirical research. Sadler (2009), as described above, presents anecdotal evidence.
	Context-dependence	71. There is debate as to whether a rubric needs to be task specific so that it applies to a single task or generic so that the same rubric can be applied to a number of different tasks (Popham, 1997; Wiliam, 2011). The debate emanates from the desire for rubrics to have broader applicability and thereby to help students generalize learning from one context to another. “Rubrics are often used by teachers to grade student work but many authors argue that they can serve another, more important, role as well: When used by students as part of a formative assessment of their works in progress, rubrics can teach as well as evaluate” (Reddy & Andrade, 2010, p. 437). (p. 254)	This is not really a critique, but an exposition of facts about the tension between deciding for a task specific or general rubric.

	Standardization and narrow the curriculum	72. Rezaei and Lovorn (2010) reported a study in which participants were asked to grade one of the two samples of writing, assuming it was written by a graduate student, once using a rubric and once without a rubric. Their results showed that the raters were significantly influenced by mechanical characteristics of the students' writing rather than the content, even when they used the rubric. This led them to ask: "if a rubric like the one used in this project, which was designed by a group of professors in a college of education, is shown to be unreliable, then what does this say about the thousands of rubrics being used every day in schools?" (Rezaei &, but Lovorn, 2010, p. 29). (p. 254)	The article by Rezaei & Lovorn (2010) provides empirical evidence, but the study contains threats to its experimental validity and such conclusions cannot be extracted from their data.
Sadler, 2014	**Limitations of criteria**	73. /…/ the qualitative and quantitative terms used in codifications of achievement standards lack the necessary linguistic properties to carry true standards successfully /…/ (p. 284)	Theoretical contribution. Occasional references are made to other fields of study, such as philosophy (Wittgenstein) and cognitive psychology (Abercrombie), but mostly to own previous publications.
Ito, 2015	**Limitations of criteria**	74. Bloxham et al. (2011), for example, argue that the majority of teachers do not use written rubrics in their marking. (p. 34)	Bloxham et al. (2011) provides empirical evidence, reporting that tutors did not make use of written criteria in their marking. It should be noted, however, that the sample is small (i.e. 12 lecturers) and that the findings therefore cannot be used to support any general claims, such as "the majority of teachers", beyond the actual sample.
		75. The literature review suggests that instructors do not use written rubrics but rather mental grading mechanisms. Teachers may use multiple levels (first, upper second, lower second, third, fail / in the 50s, upper 40s, lower 40s / A, B, C, F), or alternatively use only one dimension (see Figure 3). For instance, the respondent in the study by Bloxham et al., whose response is detailed above, did not mention any dimension. (p. 35)	An ambiguous statement, where it is not clear which studies the author might be referring to. As mentioned above, the sample size and design of the study by Bloxham et al. cannot be used to support any claims about instructors in general.
	Context-dependence	76. Shay (2004) argues that assessment, including rubrics use, is a context-dependent, experience-based and situational judgment. (p. 38)	Shay (2004) does not present any empirical evidence for the claim maintained by Ito. However, Shay does cite another author (i.e. Broad, 2000), who provides empirical evidence, which means that Ito is incorrectly citing Shay as the source. Finally, Shay critique rubrics, citing others' work on: (1) not eliminating disagreements, (2) rubrics being silent to "prejudices and foreknowledge" that are inevitable and valuable in interpreting a text, and (3) rubrics being silent about the value system underlying interpretative acts. Points (2) and (3) may be considered to lend support to the critique regarding the context-dependence of rubrics.
	Miscelanea	77. Some scholars argue that rubrics are time-consuming. Wolf and Stevens (2007) state that creating rubrics, "especially writing the descriptions of performances at each level" is time-consuming and thus "should be developed for only the most important and complex assignments" (p. 13). In Reynolds-Keefer's (2010) study, one respondent reported that making and/or using rubrics "seems really complicated…you have to know too much stuff ahead of time. It is easier to just grade" (p. 1). Another simply said, "I think it would take too much time, and I don't know how I decide how many points everything is worth" (p. 1). This time-consuming process can be stressful for both instructors and students. (p. 38)	Wolf & Stevens (2007) is a theoretical paper presenting rubrics providing only anecdotal evidence. The author's presentation of Reynolds-Keefer is not complete. The two students mentioned are from a group of 16 respondents (out of 43) who reported not planning to use rubrics in the future. This is in contrast to the 27 that planned to use them, which might be a sort of confirmation bias.

	Standardization and narrow the curriculum	78. Some scholars are concerned that rubrics could undermine, constrain, and diminish creativity (Wolf and Stevens, 2007). Linda Mabry (2013), for example, argues that rubrics may help students obtain higher scores but may also produce 'vacuous writing' (p. 678). Bloxham et al. (2011) warn that rubrics can mislead students (and teachers) that there is "something fixed, accessible and rational that they can use to guide their work (p. 663*)". (p. 38) **According to our pdf this should be page 658.*	Wolf & Stevens (2007) is a theoretical paper providing only anecdotal evidence. Mabry (2013) is likely to be an incorrect citation, as in the list of references Ito included Mabry (1999). As presented earlier, this publication is largely based on anecdotal evidence. The reference to Bloxham et al. (2011) is not made in relation to the empirical findings from this study, but in relation to the introduction. Bloxham et al. make reference to Shay (2004) and Sadler (2009), both of which are discussed above.
	Limitations of criteria	79. The findings show that many of the instructors in the sample were unfamiliar with rubrics. Some of those who knew about rubrics did not use them for specific reasons. These included that they require too much time and effort. As one respondent claimed, some instructors teach large numbers of students and grade hundreds of reports at a time. Also, rubrics are technically difficult to use. One respondent reported, for example, that the difference between level 3 and 4 of a certain item is often judged subjectively and thus inconsistently. (p. 43)	Empirical research. These are the empirical findings reported by the author, based on interviews and a focus-group. However, the interpretations have validity threats (e.g. most of the participants had not heard of rubrics, but they still provided comments on the effects of using rubrics).
Jones, Allen, Dunn, & Brooker, 2016	**Limitations of criteria**	80. For example, markers may allocate a final grade for each category on the rubric by circling a grade (Stevens & Levi, 2013) based on a subjective or comparative judgement and do not account for variation in the quality of a student's work within a given criterion (Saddler & Andrade, 2004). Used this way, the rubric lacks constructive feedback, is potentially vague and confusing, and neither fully informs students nor assists them to recognise the strengths and weaknesses in their work (Stiggins, 2001). (p. 3)	Stevens & Levi (2013) is used by Jones et al. to indicate that rubrics can be used to allocate a final grade, which is definitional, so there is no need for empirical evidence. Saddler & Andrade (2004) is an information piece and the authors did not state anything similar to what Jones et al. claim. Stiggins (2001) is a book on assessment mostly directed towards primary and secondary teachers, including practical guidelines on how to construct different types of items and tests.
	Simple implementations don't work	81. If students do not understand the rubric terminology and cannot differentiate between academic standards, rubrics have little value for either preparation or feedback. Terms such as 'critically analyse', 'synthesise' and 'evaluate' are often unclear to students and need further explanation (Chanock, 2000; Handley, den Outer, & Price, 2013). (p. 3)	The first sentence, which is the one including the term rubric, does not have any references. Then the authors used two references to back up that students have difficulties understanding the differences between those different cognitive actions, which is logical as this can prove difficult. Therefore there is no empirical evidence against rubrics in this excerpt as neither Chanock (2000) nor Handley et al. (2013) make references to rubrics in their articles.
		82. Although students may understand the terminology and standards used in rubrics, they may not be able to use criteria to develop the quality of their work and improve their performance (Langan & Wheater, 2003; Rust et al., 2003). (p. 3)	Langan & Wheater (2003) present two studies on the implementation of peer assessment. They do not mention the term rubric. Rust et al. (2003) provide empirical evidence, but does not support the claim made by Jones et al. First, Rust et al. used a "marking criteria grid", which is not necessarily the same thing as a rubric. As the marking grid was not included, no conclusions can be drawn regarding this point. Assuming that the marking grid shared the characteristics of a rubric, all participants used the sheet, and all seem to have improved their performance (p. 161). Therefore, the claim cannot be substantiated based on the findings reported by

			Rust et al. Furthermore, the workshop participants received different interventions and it is impossible to disentangle what caused the differential effects. It should be noted that this is the report of an intervention, as presented by the authors, not an experimental design study.
		83. Although students may understand the terminology and standards used in a rubric, they may not be able to use the rubric to create and evaluate their own assessment. (p. 3)	No empirical evidence for the claim.
Wollens-chläger, Hattie, Machts, Möller & Harms, 2016	**Simple implementations don't work**	84. In their review, Panadero and Jonsson (2013) conclude that rubrics make learning goals explicit and, in turn, this transparency leads to improved student performance and self-regulation processes. However, findings concerning this explanation are inconsistent. In some studies where rubrics have only been handed out to students, i.e., without any specific training (e.g., Duke, 2003; Toth, Suthers, & Lesgold, 2002) or without any further feedback information (Wollenschläger, Möller, & Harms, 2011; 2012), no significant effects on students' performance, motivation, or self-regulation could be found, although learning goals and assessment criteria were transparent to the students. Thus, it seems that transparency is not enough to make rubric feedback effective. Therefore, the first question of this study is: Is it really the transparency of learning goals that makes rubric feedback effective? Or is there another factor that is crucial to rubric feedback effectiveness on students' performance, motivation and self-regulation processes? (p. 1)	Duke (2003) compared a group using a rubric and explanations of the criteria against a group that was also trained on cognitive strategies. Therefore, the claim is not maintained by this study, as there was no control group to compare the "only rubric" condition. Toth et al. (2002) used a 2x2 research design in which rubrics were always accompanied by another instructional intervention, and all conditions performed peer reviews. Therefore, the claim is not maintained by this study, as there was no control group to compare the "only rubric" condition. The two other papers are in German. Importantly, Wollenschläger et al. is an empirical study in itself testing the theme. However, they did not include a control group not using rubrics. The study has three conditions: (1) just handing out the rubric, (2) rubric + individual performance-information, and (3) rubrics + individual performance-improvement-information. Therefore, they can only conclude that a more formative and complex implementation of rubrics can produce better results. They cannot, however, conclude that just handing out the rubric does not have an effect, because of the lack of control group.
Dawson, 2017	This article does not critique rubrics. "Just" presents the taxonomy.		
Hudson, Bloxham den Outer & Price, 2017	**Limitations of criteria**	85. The article identifies a significant gap between the theoretical positions asserted in the research literature and the conceptions held by experienced academics tasked with guaranteeing national standards. It considers implications for quality assurance and reflects on whether the dominance of transparency and accountability discourses leads academics to contort the way they talk about standards. (p. 1309)	Empirical research. 24 external examiners' responses to interview questions about their examining practice: (a) Examiners frequently expressed beliefs that relying entirely on explicit and traceable processes for judging student work is most fair, objective and transparent; (b) The form of fairness the examiners expressed most concern for was fairness to the students in terms of marking to the criteria the students had been given; (c) The concept of standards as tacit and contextual, requiring constant, situated negotiation was marginal in the interviews.
Bearman & Ajjawi, 2018	**Limitations of criteria**	86. One of the most common challenges in writing assessment criteria is capturing holistic tacit knowledge; and many argue that this knowledge is impossible to capture explicitly (O'Donovan	Conceptual/theoretical paper making references to both empirical research and theoretical contributions. O'Donovan et al. (2004), analyzed above, report on a "marking workshop" intervention, where findings (replicated for 3 years)

		et al., 2004; Orr, 2007; Sadler, 2009; Bloxham and Boyd, 2012; Bloxham et al., 2016; Hudson et al., 2017). (p. 3)	show that students who undertake the workshop demonstrate a significant improvement in performance compared with those who do not. Orr (2007) is a mostly theoretical contribution, although some selected empirical findings are reported. Details on methodology are rudimentary. Sadler (2009), analyzed above, is a theoretical contribution, based mainly on anecdotal evidence. Bloxham and Boyd (2012) is based on the same empirical data as Bloxham et al. (2011), analyzed above, which means 12 lecturers from two universities asked to "think aloud" as they graded two written assignments. Findings suggest that tutors believe there are established and shared academic standards in existence for their discipline and they try to uphold them. Bloxham et al. (2016) used a repertory grid methodology to examine the implicit assessment criteria among 24 assessors in four different disciplines. Findings show variation in the choice, ranking and scoring of criteria. Authors claim that assessment decisions are so complex, intuitive and tacit that variability is inevitable. Hudson et al. (2017), analyzed above, report on 24 external examiners' responses to interview questions about their examining practice. The references provide support for the claim that capturing holistic tacit knowledge is challenging and there are authors who claim that capturing such knowledge is impossible, although the limited data hardly warrant such strong generalizations.
	Context-dependence	87. /…/ if academic standards are socially constructed and based on tacit, dynamic knowledge, then how these standards are perceived and how the knowledge is understood, depends on an individual's social history and standing. This applies equally to assessment criteria. (p. 3)	References are not seen in the excerpt, but are made to: Jankowski & Provezis (2014), which is a theoretical contribution on neoliberal ideologies, governmentality and the academy; and Tummons (2014), which is a theoretical contribution problematizing professional standards for teachers in the UK lifelong learning sector. Both references are scientific, but non-empirical, and represent a view that is in line with the claim made by Bearman & Ajjawi.
	Limitations of criteria	88. The text of any written assessment criteria suggests that particular forms of knowledge are particularly important: students should be paying attention to *this*. In doing so, without mentioning it, they direct students' attention away from *that*. So the whole notion of transparency must inevitably be based on highlighting some things and *obscuring others*. (p. 3)	References are not seen in the excerpt, but is made to Strathern (2000), which is a theoretical contribution presenting a social anthropologist view on accountability in education. The reference is scientific, but non-empirical, and represent a view that is in line with the claim made by Bearman & Ajjawi.

	Miscelanea	89. /…/ the very notion of transparency can be seen as contributing to a system that seeks to commodify and control education (Brancaleone & O'Brien, 2011). From this view, transparency enacted through written assessment criteria permit institutions to enforce governance. By ensuring assessment criteria are visible, institutions have a means to control teachers and teaching. (p. 3)	Brancaleone & O'Brien (2011) is a theoretical contribution on the influence of marketisation on education. The reference is scientific, but non-empirical, and represent a view that is in line with the claim made by Bearman & Ajjawi.
	Instrumentalism and "criteria compliance"	90. Many students seek to use the written criteria to pass the assessment rather than learn (Norton, 2004; Bell et al., 2013). Academics are all familiar with students who come, checklist in hand, saying 'why did I get this mark?' (p. 4)	Norton (2004) is a case study in psychology. The author argues that explicit criteria may have a deleterious effect on students' learning. The empirical grounding for that argument is, however, quite weak and do not lend itself to generalization beyond the specific sample: "Additional guidance on what the assessment criteria meant was provided in the module handbook. Unfortunately, adding more detail about the assessment criteria in this way had the unintended effect of apparently raising students' anxieties and leading them to focus on sometimes quite trivial issues. /…/ This experience has challenged the author's assumptions that a focus on making assessment criteria explicit will enable students to take a deep approach." (p. 693) Bell et al. (2013), analyzed above, report on first-year students' perceptions of the usefulness of grade descriptors, marking criteria and annotated exemplars. Of 119 students, 87% found the resources to be useful. However, students differed in terms of seeking more precise guidance or being content with a general idea of the standards. The references support to the claim that there are students who seek to use the written criteria to pass the assessment rather than learn [course content]. Whether the references support any generalizations regarding the number or share of students who display this disposition, or under which circumstances, is doubtful.
		91. Our discourses of transparency may also affect how students see knowledge. Saying that our assessment criteria are 'transparent' reveals how we understand knowledge itself, or our epistemic beliefs, and therefore power. From an educational perspective, what we want to avoid, is giving our students the sense that knowledge is fixed and stable. (p. 4)	References are not seen in the excerpt, but is made to Ajjawi and Bearman (2018), which is a theoretical contribution on different perspectives on standards. The reference is scientific, but non-empirical, and represent a view that is in line with the claim made by Bearman & Ajjawi. It does not, however, explicitly support the claim that discourses of transparency affect how students view knowledge.

Bouwer, Lesterhuis, Bonne & De Maeyer, 2018	**Standardization and narrow the curriculum**	92. Results show that the instructional approach [i.e. comparative judgment vs. the use of rubrics] influenced the kind of aspects students commented on when giving feedback. Students in the comparative judgment condition provided relatively more feedback on higher order aspects such as the content and structure of the text than students in the criteria condition. This was only the case for improvement feedback; for feedback on strengths there were no significant differences. (p. 1)	Empirical research. This research investigated the learning effects of two different instructional approaches: applying criteria to examples and comparative judgment. International business students were instructed to write a five-paragraph essay, preceded by a 30-min peer assessment in which they evaluated the quality of a range of example essays. Half of the students evaluated the quality of the example essays using a list of teacher-designed criteria (criteria condition; n = 20), the other group evaluated by pairwise comparisons (comparative judgment condition; n = 20). Students were also requested to provide peer feedback. /.../ Positive effects of comparative judgment on students' own writing performance were only moderate and non-significant in this small sample.
Brookhart, 2018	**Context-dependence**	93. Sadler (2014) argued that codification of qualities of good work into criteria cannot mean the same thing in all contexts and cannot be specific enough to guide student thinking. He suggests instantiation instead of codification, describing a process of induction where the qualities of good work are inferred from a body of work samples. (p. 2)	Sadler (2014), analyzed above, is a theoretical contribution with occasional references to other fields of study, but mostly to own previous publications.