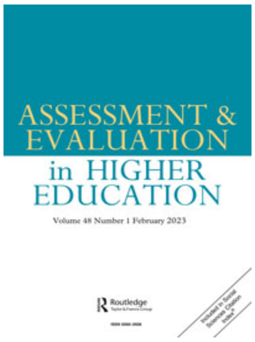
ASSESSMENT & EVALUATION in HIGHER EDUCATION
Volume 48 Number 1 February 2023

Routledge
Taylor & Francis Group

Assessment & Evaluation in Higher Education

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/caeh20

Supervisors’ untrained postgraduate rubric use for formative and summative purposes

Lieselotte Postmes, Rianne Bouwmeester, Renske de Kleijn & Marieke van der Schaaf

To cite this article: Lieselotte Postmes, Rianne Bouwmeester, Renske de Kleijn & Marieke van der Schaaf (2023) Supervisors’ untrained postgraduate rubric use for formative and summative purposes, Assessment & Evaluation in Higher Education, 48:1, 41-55, DOI: 10.1080/02602938.2021.2021390

To link to this article: https://doi.org/10.1080/02602938.2021.2021390

Published online: 06 Jan 2022.

ASSESSMENT & EVALUATION IN HIGHER EDUCATION
2023, VOL. 48, NO. 1, 41–55
https://doi.org/10.1080/02602938.2021.2021390

Routledge
Taylor & Francis Group

Supervisors' untrained postgraduate rubric use for formative and summative purposes

Lieselotte Postmes[a], Rianne Bouwmeester[b], Renske de Kleijn[b] and Marieke van der Schaaf[a]

[a]Utrecht Center for Research and Development of HPE, University Medical Center Utrecht, Utrecht, The Netherlands; [b]Biomedical Sciences, University Medical Center Utrecht, The Netherlands

ABSTRACT

Using rubrics can benefit the quality of assessment and learning. However, the conditions that stimulate or obstruct these benefits have been insufficiently studied. One underinvestigated claim is that rubrics are no substitution for good instruction and assessment and that teachers need training in utilising them. This is relevant since teachers in daily practice often use rubrics without training. In this study, we investigated data from a rubric filled out by supervisors who were not specifically trained in its use and used the rubric voluntarily. The rubric was designed for assessment moments with a formative and summative purpose. Results of quantitative analyses of 313 rubric forms indicated that the rubric was used flexibly: supervisors vary in using the rubric for the formative and/or summative purpose and in the criteria they assess. More criteria were omitted during formative use than during summative use. Some of these omitted criteria were most predictive for the final grade. This raises serious concerns with respect to the tension between flexible rubric use and constructive alignment. To understand the quality of rubric use in education, future research is needed on supervisors' perceptions toward rubrics use.

KEYWORDS
Rubrics; formative assessment; summative assessment; research skills

Introduction

Rubrics as assessment tools that articulate the criteria and standards for students' work are often used in higher education (Goodrich 1997; Popham 1997; Dawson 2017). Higher education research performed in the past decades shows a considerable interest in rubrics (Jonsson and Svingby 2007; Reddy and Andrade 2010; Panadero and Jonsson 2013; Brookhart 2018; Panadero and Jonsson 2020). A large part of this research shows potential benefits of rubrics use from a student and teacher perspective. Students' rubric use can positively influence student learning, improve performance and self-regulation. Regarding teacher use, rubrics can contribute to the quality of assessment that improves students' learning and especially reliability. However, as Brookhart (2018) and Panadero and Jonsson (2020) pointed out, rubrics are not automatically beneficial. The attention in research is shifting to the importance of the circumstances that attain or limit these benefits (Brookhart 2018; Panadero and Jonsson 2020).

CONTACT Lieselotte Postmes l.m.l.postmes-2@umcutrecht.nl

In their narrative review on rubric critiques, one of the themes Panadero and Jonsson identified is that "*simple implementations of rubrics do not automatically work*". The main underlying criticisms of this theme are: (1) rubrics are no substitution for good instruction and assessment, (2) simple interventions such as just handing out rubrics are not enough, and (3) teachers need training (Panadero and Jonsson 2020). They found supporting literature for the second claim that simple interventions do not work, but they could not empirically support the first and third claims. To optimise rubric use, it is relevant to investigate whether teachers need training and what the consequences are if this training is missing, especially since "simple" rubric implementations in which teachers are not trained are common in daily practice. Therefore, this study aims to further explore how teachers use a rubric that they are not specifically trained for. We focus on supervisors' rubric use in formative and summative assessment in the context of students' research skills during research internships in their post-graduate biomedical sciences education. This setting is chosen as it is an exemplar of teachers' rubrics use by supervisors in post-graduate education in the life sciences worldwide.

Theoretical framework

Rubrics articulate expectations for student work by listing criteria for the work and performance level descriptions across a continuum of quality (Brookhart 2018). More specifically, rubrics can help to assess the quality of students' performance of specific tasks by providing relevant criteria and information on the corresponding quality descriptions. Often table-shaped, the rows represent the criteria and the columns represent the various quality levels. Each cell contains a description of a particular quality level. This way, the criteria for the tasks and the desired quality become explicit. An assessor, whether students themselves, peers or teachers, can indicate where within this table an assessee's performance is currently located and thus provide both feedback and feedforward (Bradley, Anderson, and Eagle 2020).

Rubrics can be used for formative and summative assessment purposes (Panadero and Jonsson 2013; Brookhart and Chen 2015). Summative use of rubrics refers to applications that primarily aim at high stake decisions of assessment, evaluation and accreditation processes. Formative rubric use primarily aims at supporting student learning. This is supported, for example, by guiding teachers in their feedback provision, making the qualities of good work explicit and/or providing students with insight into their current performance and the demands of their expected performance (Brookhart and Chen 2015; Prins, de Kleijn and van Tartwijk 2017).

Given the differences in impact for students' learning, the design of rubrics with a formative or summative purpose can vary. Rubrics used for a formative purpose can be more extensive and flexible since there is less need for standardisation than when used for summative purposes (Panadero and Jonsson 2020). An example of flexibility is that teachers can adapt or even design rubrics themselves for their specific use, possibly collaborating with their students or colleagues (Kilgour et al. 2020). However, the literature offers no advice on what design to adopt if a rubric is used for both formative and summative purposes. From a constructive alignment point of view (Cohen 1987; Biggs 1996), it seems desirable to keep the design similar for both purposes. In this way, the attention of both assessor and assessee is directed to the criteria that are important for the summative assessment and the feedback provided during the formative moment(s) can be directed and focused accordingly.

Venning and Buisman-Pijlman (2013) describe an 'assessment matrix' (a rubric) used to promote research skill development in postgraduate research projects. They conclude that their matrix could be well integrated into both formative and summative assessments. Their study is one of the few on rubrics for both summative and formative purposes, focusing on research skills and rubrics within the postgraduate context. A specific feature of assessing students' research projects in postgraduate environments, for instance during research internships, is that assessments often require flexibility and adaptivity, as research projects within one program are generally not uniform for all students. One of the positive feedback strategies identified by Chugh, Macht, and Harreveld (2021) is to let supervisors provide timely and constructive feedback e.g. via regular meetings. The current study in postgraduate education investigates a rubric that supports timely and constructive feedback by having both a formative and summative purpose.

Most literature on formative rubric purposes focuses on how (undergraduate) students use rubrics, for instance, in peer feedback situations (Panadero and Jonsson 2013). So far, little is known about how teachers use rubrics during their assessments of students' work, particularly when it comes to formative purposes. Research on teachers' rubric use in postgraduate education is scarce (Brookhart 2018), but studies in other educational settings describe how teachers use of rubrics may influence the focus of their feedback process. Jeong (2015) described how university teachers' feedback can differ with and without using a rubric and how the rubric can make teachers' focus on different assessment elements. A study by Kutlu, Yildirim, and Bilican (2010) indicated that primary school teachers with positive attitudes to rubrics seemed to benefit more from their use, needed to consult colleagues less often and were more prone to use the rubric for feedback as opposed to grading. More research on this topic is important since, compared to student feedback, teacher's formative rubric feedback is more valued by students, contains more extensive and specific comments in areas where expertise is required, and can have a greater impact on student learning (Hamer et al. 2015; van Ginkel et al. 2017).

The majority of teacher assessment literature on rubrics has focused on summative assessment and achieving high(er) reliability in scoring (e.g. Oakleaf 2009; Rezaei and Lovorn 2010; Park et al. 2016; Menéndez-Varela and Gregori-Giralt 2018). In the literature, it is often advised to use rater training to ameliorate inconsistencies in the scoring process of teachers (Reddy and Andrade 2010). However, in line with the conclusion of Panadero and Jonsson (2020) that simple implementations (such as just handing out rubrics) do not automatically work, the empirical literature seems to indicate that short and easily accessible digital training courses were found to have little impact on inter-rater agreement (Pufpaff, Clarke, and Jones 2015; Davis 2016). Increasing the duration and training intensity of any type of training does seem to benefit the interrater reliability (Lovorn and Rezaei 2011; Davis 2016), and what improves the reliability most, is the addition of a consensus training (Shafer et al. 2001; Lovorn and Rezaei 2011). In other words, only very time-consuming rater trainings seem to benefit the reliability of rubric scoring. This is where scientific findings conflict with daily practice, as such trainings are commonly not feasible in practice, given time and practical constraints (Oakleaf 2009; Pufpaff, Clarke, and Jones 2015; Broadbent, Panadero, and Boud 2018).

In summary, studies on summative rubric use mainly focus on reliability and the importance of teacher training and studies on formative rubric use focus on student use. This study aims to contribute to the knowledge base on rubric use by studying how teachers, without explicit rubric training, use a rubric when they can use it for both formative and summative purposes. The significance lies in the shift of focus from reliability and teacher training to untrained rubric use that is common in daily practice. Thus, the purpose of this study is to contribute to filling this gap by answering the following research question: *How do supervisors with no specific rubric training employ a rubric for research skills in an assessment with formative and summative purposes?*

Present study

Educational context

In our educational context, students' development of research skills is essential. Therefore we focused our study on the 'research skill rubric' used during research internships in Dutch biomedical science graduate education at Utrecht University in the Netherlands. The context of the study were two student research internships with a duration of 6 and 9 months. During these internships students join a research group to perform experimental, biomedical research at a university or institution, in the Netherlands or abroad. Students were supervised by two teachers: a daily supervisor on location and an examiner affiliated with our graduate school. The research skills on which we focus in this study were solely assessed by the daily supervisor. During the research internships, there were two mandatory assessment moments with an interval of 2 to 6 months in which the research skills were discussed. The first assessment moment, halfway through the internship, was the interim assessment meant to provide formative feedback to the students. The second, at the end of the internship, was the final assessment which was summative in nature since a grade is given.

The assessment moments' purposes (formative or summative) were communicated beforehand to both supervisors and students. In the past, supervisors had indicated unfamiliarity with the expected level of student performance. Therefore rubrics had been introduced as a way to provide general criteria and standards. The use of rubrics was made voluntary to reduce expected resistance by supervisors in the implementation process. This voluntary nature of using rubrics aligns with Kutlu, Yildirim, and Bilican (2010) findings that teachers with positive attitudes to rubrics seem to benefit more from their use. As an alternative to the rubric, supervisors could provide a written report of the interim assessment meeting and a written justification of the summative grade.

Daily supervisor context

Although all daily, or local, supervisors were experts in the field of biomedical sciences and approved as supervisor by the board of examiners, their background varied. Supervisors came from different subspecialities within biomedical sciences and represented a wide variety of institutions and universities that students could go to for their internships. They were generally members of a research group, mostly staff or sometimes non-staff members (i.e. PhD student or post-doc). Given this variety, we expect their prior experiences with rubric assessment and attitudes towards rubrics to differ. Before each mandatory assessment moment, supervisors received e-mails containing links to information on learning objectives, a research project guide, the interim assessment and the voluntary rubrics. No teacher training for any of the assessments, including using the research skill rubric, was introduced since it was deemed infeasible in the light of daily supervisors' time restraints and distributed locations.

The graduate school of life sciences' research skill rubric

The research skills rubric had been developed by the board of examiners of the Graduate School of Life Sciences (GSLS), who monitor the quality of the research internship. The rubric criteria were based on the learning objectives of the internship and further developed and refined during several pilots. To enhance validity, this was done in close cooperation with a dozen program coordinators and about twenty examiners, who provided feedback regularly via meetings and focus groups. This resulted in a generic rubric that was used for institute-wide assessment of research skills. The reliability of the scores that supervisors gave, based on the rubric, could not be estimated because of the practical constraint of having only one daily supervisor to supervise and assess the student on a regular basis. Furthermore, teacher training was deemed impractical, and the board of examiners felt that the rubric could provide valuable feedback for students, regardless its reliability.

The rubric consists of 13 evaluative criteria that are classified into three main categories: performing research (PR), practical skills (PS) and professional attitude (PA). Several criteria consist of multiple sub-criteria. Each (sub)criterion generally consists of quality descriptors for three quality levels (insufficient, satisfactory and excellent). This seemed to provide the most meaningful quality descriptors. However, some criteria deviated, in accordance with the advice to relate the number of quality levels to the type of decision and the number of reliable distinctions (Brookhart 2018). The best example of this is the 'integrity & conscientiousness' (PA) criterion which was viewed as a yes/no decision. This resulted in a quality descriptor for the insufficient (not integer) and satisfactory (integer) quality levels; the box for the quality level of 'excellent' therefore contains no description. The same rubric can be used both during the (formative) interim assessment and the (summative) final assessment. When the rubric is used during the interim assessment for formative purposes, it is a means to facilitate supervisors clarifying their expectations and providing feedback. It is also a means for students to help them plan and self-assess. When the rubric is used during the final assessment with a summative purpose, it is meant to support supervisors' holistic grading. In that case, supervisors use the criteria that are relevant to them and aggregate these into an overall judgement represented in a single grade. No formula was used to calculate a final grade. The rubric itself can be found in Appendix 1.

Methods

Data collection

Data from January 2015 to May 2018 was provided by the Administration Office of the biomedical science faculty of the GSLS for nine research master programmes. The Administration Office is responsible for collecting assessment forms, such as the research skill rubric forms. Of these submitted forms, we collected the data on the research skill grade, the moment of rubric use (formatively during the internship or summatively at the end of the internship) and the quality levels of the rated evaluative criteria. The traceable student number was anonymised. During this period, 980 research internships were completed. We collected 313 research skill rubrics from 237 internships (24.2% of total amount).

Procedure

The research skill rubric has three quality levels (insufficient, sufficient and excellent) and 13 criteria, of which eight consist of multiple sub-criteria (see Appendix 1). When rating a (sub) criterion, supervisors are allowed to tick multiple quality levels if they deem the students' level between insufficient/satisfactory or between satisfactory/excellent. To allow quantitative evaluation required for this study, we recoded supervisors ratings into five scores (1.0, 1.5, 2.0, 2.5 and 3.0). A 1.0 was assigned if the criterion was (fully) judged as insufficient, a 2.0 for satisfactory and a 3.0 if deemed excellent. A 1.5 was assigned when both insufficient and satisfactory levels were scored, either in different sub-criteria or in one criterion. The same would apply to 2.5, only then if the supervisor scored between the levels of satisfactory and excellent.

Analysis

We analysed the recoded scores of the completed rubric forms to gain insight into how supervisors who had received no specific rubric training employ a rubric for research skills in their assessment moments with formative and summative purposes (during and at the end of the internship). First, absolute amounts and percentages of the completed rubrics were mapped to calculate how often the rubric is used for formative and summative purposes. For the summative assessment purpose, we calculated the average of the research skill grade and the criteria scores' estimated reliability using Cronbach's alpha. Next, criteria use per assessment purpose was further investigated by calculating range, mean, standard deviations (SD) and missing data of the recoded scores. Rubric completeness was further explored by mapping missing data per criterion as the total amount per rubric. Finally, to determine which rubric criteria are most predictive, a multiple regression on the summative feedback data was conducted to explore to what degree criteria predict the final grade. A forced entry method was used to prevent random variation in the data that can be observed in stepwise techniques. Cases with missing values were deleted listwise.

Ethical considerations

The Netherlands Association for Medical Education Ethics Review Board approved the study (case number 2019.2.4). No participant consent was required since we only used anonymised data.

Results

Formative and summative rubric use

The collected data consists of 313 research skill rubric forms containing information from 237 internships. The rubric was mostly used solely during the final assessment that had a summative purpose ($n = 102$, 43.0%). In almost a third of the internships ($n = 76$, 32.1%), supervisors used the rubric for both the interim and final assessment moments, for both formative and summative use. The rubric was least often used solely to provide formative feedback during the interim assessment ($n = 59$, 24.9%). So, use during the interim assessment did not guarantee use during the final assessment: in almost half of the internships in which the rubric was used for a formative purpose, the rubric was not subsequently used for a summative purpose (59 of 135, 43.7%: see Figure 1).

The research skill grade, provided for the 178 summative rubrics, had an average of 8.24 (SD = .79) and ranged from 6.0 to 10.0. The criteria scores for the summative assessment had an estimated reliability of Cronbach's alpha .92, indicating high internal consistency among scores on items.

Criteria used in formative and summative feedback

The descriptives of scores on the 13 criteria are presented in Table 1. All criteria but three were scored within the range of available quality levels. Noteworthy is the criterion of 'integrity &

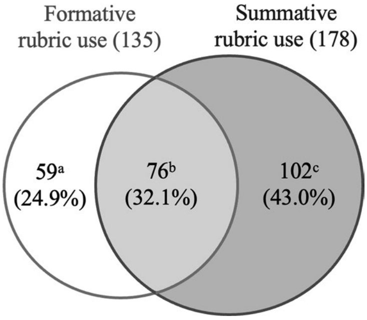

Figure 1. Distribution of internship rubric use.a. 59 internships with solely formative rubric use (59 rubric forms); b. 76 internships with both formative and summative rubric use (resulting in 152 rubric forms); c. 102 internships with solely summative rubric use (59 rubric forms)

Table 1. Descriptive statistics for criterion scores used and omitted in the rubric assessment of research skills for formative and summative use

Rubric criterion	Formative rubric use						Summative rubric use					
	Completed criteria				Omitted		Completed criteria				Omitted	
	n	Range	Mean	SD	n	(%)	n	Range	Mean	SD	n	(%)
Performing Research (PR)												
Design research	127	1.0 - 3.0	2.11	.50	8	5.9	175	1.0 - 3.0	2.27	.54	3	1.7
Data analysis and interpretation	113	1.0 - 3.0	2.28	.51	22	16.3	178	1.0 - 3.0	2.35	.50	0	0.0
Discussing outcomes	129	1.0 - 3.0	2.23	.42	6	4.4	177	1.0 - 3.0	2.33	.50	1	0.6
Practical Skills (PS)												
Technical skills	132	1.0 - 3.0	2.32	.43	3	2.2	178	1.0 - 3.0	2.51	.44	0	0.0
Efficiency	121	1.0 - 3.0	2.36	.56	14	10.4	171	1.0 - 3.0	2.51	.55	7	3.9
Organization work records	128	1.0 - 3.0	2.20	.39	7	5.2	172	1.0 - 3.0	2.36	.42	6	3.4
Organization working place	124	1.0 - 3.0	2.38	.41	11	8.1	171	1.0 - 3.0	2.48	.44	7	3.9
Professional attitude (PA)												
Initiative, independence, creativity & handling feedback	135	1.5 - 3.0	2.36	.37	0	0.0	178	1.0 - 3.0	2.44	.42	0	0.0
Critical attitude	127	1.0 - 3.0	2.19	.43	8	5.9	174	1.0 - 3.0	2.27	.50	4	2.2
Integrity & conscientiousness	124	2.0 - 3.0	2.04	.20	11	8.1	164	1.5 - 3.0	2.05	.22	14	7.9
Perseverance & dedication	119	1.0 - 3.0	2.32	.48	16	11.9	169	2.0 - 3.0	2.57	.48	9	5.1
Communication with colleagues	129	2.0 - 3.0	2.55	.45	6	4.4	174	1.5 - 3.0	2.63	.45	4	2.2
Timelyness	131	1.0 - 3.0	2.46	.48	4	3.0	177	1.0 - 3.0	2.54	.49	1	0.6

n: number of cases, SD: standard deviation. Total n formative use = 135, Total n summative use = 178. Total of completely filled out rubrics for formative use = 87, Total of completely filled out rubrics for summative use = 140.

conscientiousness' (PA). The rubric describes only the insufficient and sufficient quality level, so this criterion contains only two quality levels. However, the mean and range indicate that some supervisors (formative use $n=5$, summative use $n=8$) did provide the student with the quality level of 'excellent', despite there being no description available.

As shown in Table 1, supervisors took the liberty to skip or omit certain criteria. In the summative assessment, the 'integrity & conscientiousness' (PA) criterion is most often omitted with 7.9% of cases, a percentage similar to the formative assessment (8.1%). In the formative rubric, the criterion 'data analysis and interpretation' (PR) is most omitted (16.3%). However, in the summative rubric, this criterion was (one of the few criteria) always filled out. The criteria of 'perseverance & dedication' (PA) and 'efficiency' (PS) are relatively often omitted in the formative rubric use (respectively 11.9 and 10.4%) and summative rubric (5.1 and 3.9%).

As shown in Table 2, supervisors omit certain criteria in about a third of cases during formative rubric use (35.6%) and a fifth of cases (21.3%) during the summative assessment. If supervisors had not completed the entire rubric, the number of missing criteria was also higher in formative rubric use than in summative rubric use. Of the incomplete summative rubrics, 71.3% of the cases had only one missing criterion, while this was 39.6% for incomplete formative

Table 2. Number of missing criteria within the incomplete rubrics.

Number of missing criteria	Formative rubric use		Summative rubric use	
	n	%	n	%
1	19	39.6	27	71.0
2	12	25.0	5	13.2
3	5	10.4	5	13.2
4	5	10.4	1	2.6
5	4	8.3		
6	3	6.3		
total	48[a]	100	38[b]	100

a: 48 of a total of 135 rubrics for formative use were incomplete (35.6%), b: 38 of a total of 178 rubrics for summative use were incomplete (21.3%)

rubrics. The maximum of missing criteria in a single rubric was four for the summative rubrics, while up to six criteria were missing in the formative rubric.

Predictive value of criteria for the summative research skill grades

Regression analysis is used to identify which criteria are most predictive for the summative research skill grade. Scores of the 'integrity & Conscientiousness (PA)' criterion offered little variance and were, therefore, not included in the analysis. Cases with missing values ($n = 27$, 15.2%) were deleted listwise.

Six of the twelve criteria are significant predictors for the research skill grade. The model explains 71% of the variance, making it a significant prediction of the research skill grade, F(12, 138) = 39,4, p < .001 (See Table 3). These results show the importance of the professional attitude category, of which half of the criteria contribute significantly to the model.

It is noteworthy that three of these six significant criteria ('data analysis and interpretation' (PR) 'efficiency' (PS), and 'perseverance & dedication' (PA)) also have the highest percentages of missings during the formative feedback provision (Table 1).

Discussion

This study contributes to the knowledge and practice of untrained rubric use by studying supervisor use of an institute-wide rubric for both formative and summative purposes. This was examined within a postgraduate context with practical restrictions for rubric teacher training, which is in line with the reality of other educational institutions. Our findings indicate that supervisors use the rubric in three flexible ways.

The first way of flexible rubric use was that supervisors vary when it comes to the purpose for which they use the rubric. We found three groups: supervisors that used the rubric solely for summative use (most common), solely for formative use (least common) and for both. This is a relevant and possibly alarming finding, as from a constructive alignment point of view (Biggs 1996), using the same rubric for a formative and summative purpose seems desirable.

Table 3. Multiple linear regression model.

	B	SE B	β	*p*	95% CI
(Constant)	3.632	.254		.000	[3.130, 4.134]
Performing Research (PR)					
Design research*	.285	.095	.194	.003*	[.10, .47]
Data analysis and interpretation	.100	.106	.060	.348	[-.11, .31]
Discussing outcomes	.141	.102	.087	.169	[-.06, .34]
Practical Skills (PS)					
Technical skills	.051	.112	.028	.651	[-.17, .27]
Efficiency*	.206	.088	.139	.021*	[.03, .38]
Organization work records*	-.247	.110	-.130	.026*	[-.47, −.03]
Organization working place	.106	.104	.058	.309	[-.10,.31]
Professional attitude (PA)					
Initiative, independence, creativity, handling feedback*	.413	.124	.211	.001*	[.17, .66]
Critical attitude	.101	.097	.063	.298	[-.09, .31]
Perseverance & Dedication*	.225	.086	.132	.010*	[.06, .40]
Communication with colleagues	.179	.103	.100	.084	[-.02, .38]
Timelyness*	.309	.103	.191	.003*	[.11, .51]
R^2	.77				
ΔR^2	.75				

PR: Performing research; PS: Practical Skills; PA: Professional attitude.

B regression beta, *p* significance value, SE B standard error unstandardised beta, β standardised beta, 95% CI 95% confidence interval, R^2 variance explained, ΔR^2 adjusted variance explained.

* p < .05

Provided that the purpose of a rubric is to transparently communicate to students, the rubric's repeated use has the potential to convey learning goals to students and provide them with feedback on how they are managing. Venning and Buisman-Pijlman (2013) investigated use of assessment matrices (rubrics) in postgraduate research projects. From data from a questionnaire, they found that the assessment matrix (rubric) was not always applied as an integral part of a (formative) feedback process, but mostly viewed as useful for assessment and the allocation of grades. In the light of their results, it is especially interesting that almost half of the supervisors who used the rubric in their formative assessment did not subsequently use it for the summative assessment. It raises the question of why this was the case.

The second way of flexible rubric use relates to supervisor assessment, and the number of quality levels used. We found that several supervisors used a non-existent quality level. It was noticeable that the criterion consisting of only two levels ('integrity & conscientiousness' (PA)) was filled in by some supervisors at the third, empty level. The number of quality levels should relate to the type of decision and the number of reliable distinctions that are possible and helpful (Brookhart 2018). That our study found that supervisors have provided their students with the quality level of 'excellent', despite the level being empty, might indicate that supervisors - contrary to the rubric's creators - did not recognize a dichotomic yes/no decision but felt that students can exhibit behaviour reflecting excellence on this criterion. This kind of information could lead to a discussion and reconsideration of the number of quality levels for this criterion. Our analyses could also have indicated that for other criteria less than three levels were helpful for supervisors, but this was not the case.

The third way of flexible rubric use was reflected in the omission of criteria with the distinct difference that criteria were more often omitted in formative use than summative use. We conclude that these supervisors, without external guidance, applied the rubric in a more flexible way during formative rubric use and thus found some first empirical evidence that flexibility for this purpose comes naturally. This finding is in line with Panadero and Jonsson's (2020) suggestion to use rubrics for formative purposes in a flexible way. An argument can be made that this flexible and varied use might have contributed to the validity of the feedback that was given in the sense that teachers were able to adapt the rubric for their specific use and were not obliged to make up feedback on criteria they had not observed yet. This also resonates well with Chugh, Macht, and Harreveld (2021) conclusion that supervisors' flexibility and adaptability are a key strategy to achieve effective feedback. Our finding of flexible rubric use suggests that the same rubric can fulfil both a formative and a summative purpose, as long as the assessment purpose is made clear to students before the moment of assessment.

However, our study also shows a clear disadvantage of this flexible rubric use. Our most striking finding is that some of the most predictive rubric criteria for the summative assessment moment, in which students got their research skill grade, were omitted during the moment in which the rubric was used formatively. In other words, students did not receive feedback during the formative assessment on some of the criteria that affect their summative assessment and grade most. The timing of the feedback might explain this partially. For instance, the most often omitted criterion of 'data analysis and interpretation' (PR) in the formative rubric could indicate that students in this stage of their research internship had not collected data yet that could be analysed or interpreted. In contrast, criteria such as 'efficiency' (PS) and 'perseverance & dedication' (PA) should be assessable during the first months of the research internship but were nevertheless omitted in more than a tenth of cases during the formative feedback moment. This indicates a missed opportunity to provide students with feedback on important aspects of their research skills and criteria determining their final grade. This might be a potential downside of flexible rubric use. Feasible solutions to this problem in this specific study context could be to expand the rubric instructions, visibly mark the rubric's most predictive criteria, or even make essential criteria mandatory.

Implications and suggestions for further research

To the best of our knowledge, this is the first study examining teachers' untrained rubric use for both formative and summative purposes. The flexible rubric use demonstrated in our study suggests that the same rubric can fulfil both a formative and a summative purpose. This is an assessment set-up and rubric design that could be of interest for others who struggle with problems such as expected resistance during the implementation process or designing a general rubric that can be applied in various contexts. Our findings show the practical application of utilising descriptive assessment statistics and regression analysis to (continuously) improve rubrics (Haagsman et al. 2021). Finally, this study could also be relevant in the development of training programs for those contexts where training *ís* feasible. Such a training could pay specific attention to criteria that are important predictors of the final grade and to the importance of providing feedback on these during the formative assessment moments.

Follow-up research could give insight into the considerations supervisors have in (not) using a rubric for formative and/or summative use when they can choose to do so. As several studies link supervisors' personal preferences to their underlying perceptions of feedback and assessment practices (Ito 2015; Chan and Luo 2021), further research in these underlying perceptions might deepen our understanding. Our study found a potential downside of flexible rubric use when criteria relevant for the final grade are omitted during formative rubric use. Others could explore which solution works best to mitigate these.

Limitations

Besides the relevant outcomes of the quantitative approach of this study, this design also brought limitations. We could not link rubric use to supervisor background information, such as (sub)specialty, experience with assessment in general, rubric assessment and the research skill rubric. As a result, we could not draw conclusions about what factors and motivations influence using the rubric (for formative and summative purposes) or the exact role of (rubric) assessment experience and training. In addition, a comment can be made about the quality of the rubric itself. Its inter-rater or intra-rater reliability has not been investigated after its development and we were unable to calculate this from the available data. Still, we did test the internal consistency of the rubric, which was high. Since our focus lies on the way the instrument is used and not the quality of the instrument itself, we think that the multiple feedback rounds, focus groups and pilots have ensured the quality of the criteria necessary to study the way the rubric and its criteria were used by the assessors. Finally, the flexible nature of the rubric contributed to more missing datapoints, i.e. criteria that were not used by supervisors in their assessment. Therefore, the multiple linear regression analysis could only be conducted on the completely filled forms. Consequently, 15% of the rubric forms was not included in our multiple regression analysis. It is possible that supervisors who omit certain criteria weigh their criteria differently.

Conclusion

This study asked, "*How do supervisors with no specific rubric training employ a rubric for research skills in an assessment with formative and summative purposes*?". We quantitatively analysed 313 research skills rubric forms for research internships in biomedical sciences. We found that the rubric was used flexibly: supervisors varied in using the rubric for a formative and/or summative purpose and in the criteria they assessed, or even omitted. To the best of our knowledge, the current quantitative study is the first to fill a gap in our knowledge on how supervisors who had received no specific rubric training employ a rubric for research skills during assessment moments with both formative and summative purposes. Our findings raise concerns with respect to the tension between flexible rubric use and constructive alignment. This warrants further research.

Notes on contributors

Lieselotte Postmes, MD, works as a Teacher and PhD candidate and her dissertation is focused on assessment and feedback in (bio)medical education.

Rianne Bouwmeester, PhD, works as an assistant professor and her research interests include (blended) innovations in biomedical education.

Renske de Kleijn, PhD, works as an assistant professor and her research interests include feedback, assessment and,research supervision, motivation, and self-regulation in higher education.

Marieke van der Schaaf, PhD, is a professor of research and development of health professions education and director of the Utrecht Center for Research and Development of Health Professions Education at UMC Utrecht, The Netherlands.

Acknowledgements

We thank Prof. Dr. Harold van Rijen, director of the GSLS in Utrecht, for enabling the data collection. We also wish to acknowledge the Assessment Panel members, from where the idea for this study sparked with particularly Dr. Gönül Dilaver.

Disclosure statement

No potential conflict of interest was reported by the authors.

ORCID

Lieselotte Postmes http://orcid.org/0000-0002-9264-8626
Rianne Bouwmeester http://orcid.org/0000-0001-9969-8207
Renske de Kleijn http://orcid.org/0000-0001-9206-4199
Marieke van der Schaaf http://orcid.org/0000-0001-6555-5320

References

Biggs, J. 1996. “Enhancing Teaching through Constructive Alignment.” *Higher Education* 32 (3): 347–364. doi:10.1007/BF00138871.

Bradley, E. J., S. Anderson, and L. Eagle. 2020. “Use of a Marking Rubric and Self-Assessment to Provide Feedforward to Level 5 Undergraduate Sport Students: Student Perceptions, Performance and Marking Efficiency.” *Journal of Learning Development in Higher Education* 2 (18): 1–23, September. doi:10.47408/jldhe.vi18.557.

Broadbent, J., E. Panadero, and D. Boud. 2018. “Implementing Summative Assessment with a Formative Flavour: A Case Study in a Large Class.” *Assessment & Evaluation in Higher Education* 43 (2): 307–322. doi:10.1080/02602938.2017.1343455.

Brookhart, S. M. 2018. “Appropriate Criteria: Key to Effective Rubrics.” *Frontiers in Education* 3 (22): 1–12 (April). doi:10.3389/feduc.2018.00022.

Brookhart, S. M., and F. Chen. 2015. “The Quality and Effectiveness of Descriptive Rubrics.” *Educational Review* 67 (3): 343–368. doi:10.1080/00131911.2014.929565.

Chan, C. K. Y., and J. Luo. 2021. “Exploring Teacher Perceptions of Different Types of ‘Feedback Practices’ in Higher Education: Implications for Teacher Feedback Literacy.” *Assessment & Evaluation in Higher Education*: 1–16. doi:10.1080/02602938.2021.1888074.

Chugh, R., S. Macht, and B. Harreveld. 2021. "Supervisory Feedback to Postgraduate Research Students: A Literature Review." *Assessment & Evaluation in Higher Education*: 1–15. doi:10.1080/02602938.2021.1955241.

Cohen, S. A. 1987. "Instructional Alignment: Searching for a Magic Bullet." *Educational Researcher* 16 (8): 16–20. doi:10.3102/0013189X016008016.

Davis, L. 2016. "The Influence of Training and Experience on Rater Performance in Scoring Spoken Language." *Language Testing* 33 (1): 117–135. doi:10.1177/0265532215582282.

Dawson, P. 2017. "Assessment Rubrics: Towards Clearer and More Replicable Design, Research and Practice." *Assessment & Evaluation in Higher Education* 42 (3): 347–360. doi:10.1080/02602938.2015.1111294.

Ginkel, S., van, J. Gulikers, H. Biemans, and M. Mulder. 2017. "The Impact of the Feedback Source on Developing Oral Presentation Competence." *Studies in Higher Education* 42 (9): 1671–1685. doi:10.1080/03075079.2015.1117064.

Goodrich, H. 1997. "Understanding Rubrics." *Educational Leadership* 54 (4): 14–17.

Haagsman, M., B. Snoek, A. Peeters, K. Scager, F. Prins, and M. van Zanten. 2021. "Examiners' Use of Rubric Criteria for Grading Bachelor Theses." *Assessment & Evaluation in Higher Education* 46 (8): 1269–1215. doi:10.1080/02602938.2020.1864287.

Hamer, J., H. Purchase, A. Luxton-Reilly, and P. Denny. 2015. "A Comparison of Peer and Tutor Feedback." *Assessment & Evaluation in Higher Education* 40 (1): 151–164. doi:10.1080/02602938.2014.893418.

Ito, H. 2015. "Is a Rubric Worth the Time and Effort? Conditions for Success." *International Journal of Learning, Teaching and Educational Research* 10 (2): 32–45.

Jeong, H. 2015. "Rubrics in the Classroom: Do Teachers Really Follow Them?" *Language Testing in Asia* 5 (1): 6. doi:10.1186/s40468-015-0013-5.

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

Kilgour, P., M. Northcote, A. Williams, and A. Kilgour. 2020. "A Plan for the Co-Construction and Collaborative Use of Rubrics for Student Learning." *Assessment & Evaluation in Higher Education* 45 (1): 140–153. doi:10.1080/02602938.2019.1614523.

Kutlu, O., O. Yildirim, and S. Bilican. 2010. "The Comparison of the Views of Teachers with Positive and Negative Attitudes towards Rubrics." *Procedia - Social and Behavioral Sciences* 9: 1566–1573. doi:10.1016/j.sbspro.2010.12.366.

Lovorn, M. G., and A. R. Rezaei. 2011. "Assessing the Assessment: Rubrics Training for Pre-Service and New in-Service Teachers." *Practical Assessment, Research and Evaluation* 16 (16): 1–10. doi:10.7275/sjt6-5k13.

Menéndez-Varela, J. L., and E. Gregori-Giralt. 2018. "The Reliability and Sources of Error of Using Rubrics-Based Assessment for Student Projects." *Assessment & Evaluation in Higher Education* 43 (3): 488–499. doi:10.1080/02602938.2017.1360838.

Oakleaf, M. 2009. "Using Rubrics to Assess Information Literacy: An Examination of Methodology and Interrater Reliability." *Journal of the American Society for Information Science and Technology* 60 (5): 969–983. doi:10.1002/asi.21030.

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

Panadero, E., and A. Jonsson. 2020. "A Critical Review of the Arguments against the Use of Rubrics." *Educational Research Review* 30 (March): 100329. doi:10.1016/j.edurev.2020.100329.

Park, Y. S., A. Hyderi, G. Bordage, K. Xing, and R. Yudkowsky. 2016. "Inter-Rater Reliability and Generalizability of Patient Note Scores Using a Scoring Rubric Based on the USMLE Step-2 CS Format." *Advances in Health Sciences Education* 21 (4): 761–773. doi:10.1007/s10459-015-9664-3.

Popham, J. W. 1997. "What's Wrong—and What's Right—with Rubrics." *Educational Leadership* 55 (2): 72–75.

Prins, F. J., R. de Kleijn, and J. van Tartwijk. 2017. "Students' Use of a Rubric for Research Theses." *Assessment & Evaluation in Higher Education* 42 (1): 128–150. doi:10.1080/02602938.2015.1085954.

Pufpaff, L. A., L. Clarke, and R. E. Jones. 2015. "The Effects of Rater Training on Inter-Rater Agreement." *Mid-Western Educational Researcher* 27 (2): 117–141.

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

Rezaei, A. R., and M. Lovorn. 2010. “Reliability and Validity of Rubrics for Assessment through Writing.” *Assessing Writing* 15 (1): 18–39. doi:10.1016/j.asw.2010.01.003.

Shafer, William D., Gwenyth Swanson, Nancy Bene, and George Newberry. 2001. “Effects of Teacher Knowledge of Rubrics on Student Achievement in Four Content Areas.” *Applied Measurement in Education* 14 (2): 151–170. doi:10.1207/S15324818AME1402_3.

Venning, J., and F. Buisman-Pijlman. 2013. “Integrating Assessment Matrices in Feedback Loops to Promote Research Skill Development in Postgraduate Research Projects.” *Assessment & Evaluation in Higher Education* 38 (5): 567–579. doi:10.1080/02602938.2012.661842.

Appendix 1. Rubric for assessing Research Skills at the Graduate School of Life Sciences

Criteria	Insufficient	Satisfactory	Excellent
Performing Research (PR)			
Design research plan / experiments	• Executes plans devised by supervisor only	• Proposes new valid experiments based on previous results • Has creative ideas	• Proposes many new, relevant experiments (with proper controls) • "Owns" the project, has original, creative ideas
Data analysis and interpretation	• Depends on supervisor for correct interpretation of results	• Provides correct analysis interpretation of results at later stages of the project	• Provides correct analysis and interpretation of results from the start of the project
	• Invalid statistical analysis	• Statistical analysis correct	• Recognizes implications
Discussing research outcomes	• Hardly participates in discussions	• Participates in discussions	• Is critical and occasionally leading during discussions
	• Fails to place research into perspective	• Discussion in the light of (recent) literature	• Stays on top of recent literature
Practical skills (PS)			
Technical skills	• Fails to master technical/lab skills • Fails to apply techniques independently	• Masters required technical/lab skills • Applies techniques independently	• Has excellent technical skills • Finds and masters new technical approach, improves existing procedures
Efficiency	• Waiting times in protocols are spent inefficiently	• Uses waiting times for preparing buffers, reading etc.	• Runs parallel experiments to use time efficiently and effectively
Organization lab journal / log/work records	• Badly organized • Required information is missing	• Well organized • All required information is available	• Repetition of experiments based on • information provided easily possible
Organization working place	• Workplace is a mess	• Tidies workplace regularly	• Workplace is always clean
	• Fails to clean equipment after use • Does not follow guidelines and protocols	• Cleans equipment after use • Follows guidelines and protocols	• Equipment is always clean • Suggests improvements for protocols
Professional attitude (PA)			
Initiative, independence, creativity, handling feedback	• Many feedback sessions are required	• Regular feedback sessions were needed	• The amount of feedback needed was minimal
	• Relies on supervisor's instructions only	• Takes initiative (initially) after stimulation	• Consults experts outside the group in consultation with supervisor, designs large parts of the project • Finds relevant new literature
	• Minimal improvement based on feedback	• Feedback led to reasonable improvement	• Response to feedback yielded excellent improvements
Critical attitude	• Critical attitude is absent • Self-reflection is absent	• Shows self-reflection and has critical attitude towards (published) research	• Critical attitude is based on intellectual depth and profundity

Criteria	Insufficient	Satisfactory	Excellent
Integrity, Conscientiousness	• Data manipulated or left out**	• Accurate, reliable and trustworthy, shows awareness of confidentiality of information	
Perseverance, Dedication	• Loses motivation when experiments / research fail(s)	• Repeats experiment until satisfactory result is obtained	• Perseveres, but knows when to stop
Communication with colleagues	• Thinks he/she is the only worker in the lab	• • Takes (needs of) colleagues into account • • Communicates with colleagues, e.g. to share equipment	• • Knows when to ask questions • • Accepts, communicates and learns from own failures
Timelyness	• Fails to meet deadlines • Fails to keep appointments	• Meets most deadlines • Keeps appointments	• Sets own deadlines and adheres to them • Schedules appointments when necessary