Author/s:
Reid, KJ;Chiavaroli, NG;Bilszta, JLC

Title:
Assessing a Capstone Research Project in Medical Training: Examiner Consistency Using Generic Versus Domain-Specific Rubrics

Date:
2022-02

Citation:
Reid, K. J., Chiavaroli, N. G. & Bilszta, J. L. C. (2022). Assessing a Capstone Research Project in Medical Training: Examiner Consistency Using Generic Versus Domain-Specific Rubrics. JOURNAL OF MEDICAL EDUCATION AND CURRICULAR DEVELOPMENT, 9, https://doi.org/10.1177/23821205221081813.

Persistent Link:
https://hdl.handle.net/11343/302328

Assessing a Capstone Research Project in Medical Training: Examiner Consistency Using Generic Versus Domain-Specific Rubrics

Journal of Medical Education and Curricular Development
Volume 9: 1–9

DOI: 10.1177/23821205221081813
SAGE

Katharine J. Reid[1], Neville G. Chiavaroli[1,2] and Justin L. C. Bilszta[1]

[1]Department of Medical Education, The University of Melbourne, Melbourne, Australia. [2]The Australian Council for Educational Research, Melbourne, Australia.

ABSTRACT: Rubrics are utilized extensively in tertiary contexts to assess student performance on written tasks; however, their use for assessment of research projects has received little attention. In particular, there is little evidence on the reliability of examiner judgements according to rubric type (general or specific) in a research context. This research examines the concordance between pairs of examiners assessing a medical student research project during a two-year period employing a generic rubric followed by a subsequent two-year implementation of task-specific rubrics. Following examiner feedback, and with consideration to the available literature, we expected the task-specific rubrics would increase the consistency of examiner judgements and reduce the need for arbitration due to discrepant marks. However, in contrast, results showed that generic rubrics provided greater consistency of examiner judgements and fewer arbitrations compared with the task-specific rubrics. These findings have practical implications for educational practise in the assessment of research projects and contribute valuable empirical evidence to inform the development and use of rubrics in medical education.

KEYWORDS: MeSH terms: Learning, research report, research, students, medical, education, medical

RECEIVED: October 15, 2021. **ACCEPTED:** February 1, 2022

TYPE: Original Research

FUNDING: The author(s) received no financial support for the research, authorship, and/or publication of this article.

DECLARATION OF CONFLICTING INTERESTS: The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

CORRESPONDING AUTHOR: Katharine Reid, Department of Medical Education, Melbourne Medical School, The University of Melbourne, Victoria, 3010, Australia.
Email: kjreid@unimelb.edu.au

Introduction

Well-designed rubrics are an important element in evaluating the quality of student performance, for both feedback and summative purposes, and can promote student engagement and self-directed learning.[1] Rubrics typically consist of three parts: relevant evaluative criteria, definitions of quality (or performance levels), and a guide to scoring the performance, either analytically (separate dimensions), holistically (an integrated judgement), or both.[1, 2] Rubrics are used extensively in medical education to assess student performance for many educational situations and performances, for both oral and written tasks. An example of the latter is the capstone research project or thesis, which usually involves complex and diverse skills in a multidimensional task, typically involving analytical, methodological and reporting components. Research projects feature in assessment in many higher degree university courses and are often used to assess research competencies in initial medical training. Nonetheless, there is little research investigating rubric development, or examiner marking performance for different forms of rubric.

Fundamental to rubric development is determining whether the criteria and quality descriptions reflect the specific task, or broader and more generalizable skills. Many rubrics focus on defining relatively precise descriptions of performance. Such descriptions are believed to promote consistent judgements between markers;[3] however, they may also inadvertently limit how generalizable the judgements are to performance contexts where similar skills are required.[2, 4] This is a key consideration in assessing student achievement for research projects where the underlying research skills may be deemed more important than specific content knowledge.[5]

Determining the reliability of a rubric, that is, the extent to which it promotes consistent marking both within a single marker and across markers is an important part of demonstrating its validity.[6–8] Validity overall, however, focuses on broader issues of appropriateness and alignment of the rubric with the purposes of the assessment and the learning context.[1, 9] Evaluating the validity of a rubric is a complex process based on multiple sources of evidence. Multiple frameworks for validation processes exist, including traditional approaches based on considerations of the content, construct and appropriate criteria.[5] More recent validity frameworks focus on the context of the assessment, in addition to validity inferences drawn from aspects such as scoring process, generalization, extrapolation and educational implications.[10,11] The first two of these, namely scoring and generaliztion, are particularly relevant to rubrics. Jonsson and Svingby[9] noted that while scoring with a rubric is likely to be more reliable than without a rubric, the same cannot be said about validity (p. 137). The validity of a rubric, they argued, is strongly influenced by its *alignment* with the relevant learning objectives, in other words, the link between skills or performance endorsed by the rubric and the pedagogical context and purpose. The design of rubrics involves thorough consideration of assessment purposes (eg assessment of learning vs. for learning), task content (eg quantitative vs. qualitative), and the stakes of specific assessments within the overall programme (eg high-stakes summative decision vs. lower-stakes continuous assessment).[12]

A key issue in evaluating the design and application of rubrics for research projects is whether generic or specific rubrics are selected.[13] Prins and colleagues characterized these differences in rubric specificity as top-down versus bottom-up approaches.[14] Top-down approaches, based on theory and expert knowledge, tend to produce rubrics with broader applicability, whereas bottom-up approaches, that use the input and expectations of teachers and students, tend to produce 'context-dependent' rubrics, closely aligned with the specific task. Yet, the research evidence for marking consistency with generic and specific rubrics is equivocal. Timmerman and colleagues[4] developed a generic (or 'universal', as they describe it) rubric (informed by content experts and the scholarly literature) to assess scientific reasoning skills in a higher education context, designed to be applicable across tasks, topics, year levels and even courses.[4] Although there was some improvement in marking reliability (which the authors attributed as much to specific and enhanced training as the rubric itself), the authors believed the major benefits were the ability to assess scientific reasoning beyond the immediate assessment context, and establish a 'common metric' for curricular evaluation. In contrast, a review of a rubric used for written case reports in a medical course, revealed that faculty preferred more specific quality descriptions, which resulted in greater inter-rater agreement.[15]

Overall, however, evidence that rubric type is related to measurement quality (particularly for research projects) is scarce. Brookhart's recent review of 46 studies describing aspects of the performance of rubrics in higher education found no relationship between rubric type and inter-rater reliability or validity.[16] Notably, only one of these studies included a review of a rubric designed to assess a research project,[14] an educational activity that may be particularly prone to poor agreement between markers.[17,18] More recently, Williams and Kemp also found low concordance between pairs of examiners for assessment of a master's thesis in psychology.[18] They also did not observe any change in marker consistency depending on whether assessment criteria were provided; however, this study did not utilize rubrics to assess the submission but relied on broad (and undefined) criteria characteristic of specific performance levels.

There is, thus, a substantial need to evaluate the measurement properties of rubrics used to assess research projects in medical training, with an emphasis on the reliability of examiner judgements for different rubric types. Such an aim accords with the perceived importance of improving the use of empirical data to guide the design and implementation of rubrics.[19] For the current study, we focussed on rubrics developed to assess a six-month research project completed by all final-year medical students in an area of interest in medicine, for which the major assessment submission is a 4000-word research thesis in the style of a journal manuscript. Our study spans four years, over which time a generic rubric was used for two years and subsequently replaced by task-specific rubrics for the following two years, with the aim to improve consistency in examiner judgements. Inconsistent marking undermines student and examiner confidence in the rigour of the examination process. Perceived support for marking consistency was a significant driver in adopting task-specific rubrics; however, empirical data were required to assess these claims. Our aim in the current study was to determine whether the consistency of examiner judgements changed dependent on whether generic or task-specific rubrics were used to evaluate the research project.

Methods

This exploratory study sought to determine whether the concordance between the marks of two examiners assigned to mark a research project varied for different rubrics by exploring whether the size of the average discrepancy in examiner marks varied by rubric type (generic vs. specific).

Current Context

The context of our study is the first semester of the final year of the Doctor of Medicine (MD) degree at the University of Melbourne, where students complete a research project that exposes students to the principles and methodologies of research and fosters experience in generating new knowledge (as opposed to simply being consumers of knowledge). Research project experience aims to develop students as clinician researchers and to foster their understanding of the intersection between research and clinical practice. To successfully complete their research thesis, students must demonstrate skills in gathering and analysing data to address a specific research question, interpreting their results, communicating their research findings, and discussing the implications of their findings in the context of existing research and knowledge. Successful completion of the subject overall also involves satisfactory completion of a literature review, progress reports, a conference poster and an overall supervisor evaluation.

Assessment of the research thesis is undertaken by two independent examiners nominated by the student's research supervisors. Examiners have expertise in the research area but are not supervising the student. For the first two years of the subject's implementation, each research thesis was assessed using a generic rubric structured according to the evaluation categories of abstract (10 marks), introduction (15 marks), methodology (10 marks), results (20 marks), discussion (25 marks), conclusion (10 marks) and organisation/presentation (10 marks) (see Appendix 1 for the generic rubric). A detailed qualitative description describing expected achievement for each of these categories was provided for each level of the university grading system (N [less than 65%], H3 [65-69], H2B [70-74], H2A [75-79], H1 [80-100]) so examiners understood the qualities of aspects of a research thesis at different

performance levels. The format of this generic rubric was based on an internal review of assessment approaches for similar academic subjects in comparable programmes.

Examiners informally identified challenges in using the generic rubric to assess the research thesis in the first two years of the subject. The generic rubric was deemed appropriate for quantitative projects, but it was perceived as difficult to apply to other types of projects (eg qualitative research or systematic reviews) because the category descriptions were inappropriate, or they lacked description of a specific project methodology. Some examiners, particularly those accustomed to marking higher degrees, continued to believe the generic rubric was too prescriptive and requested more flexibility in how marks were awarded in each category, as well as overall. In contrast, other examiners sought much more detailed analytic marking to justify a category score. Others had difficulty reconciling the maximum scores for each category against the overall grading.

Following examiner feedback the generic rubrics were reviewed, and the subsequent redesign focussed on developing task-specific rubrics. This led to the creation of six task-specific rubrics corresponding to quantitative, qualitative, mixed methods, systematic review/meta-analysis, protocol development and resource development project types (see Appendix 2 for an example of a task-specific rubric for a meta-analysis). The qualitative category descriptions and university grading descriptions were expressed as a series of questions for each category. Examiners rated each question on a five-point global scale (where 1 = Fail and 5 = Excellent). For example, for the introduction, examiners rated whether the introduction was highly focussed, whether the hypotheses and aims were clearly stated, whether the aims were clearly stated, and whether there were clear links between the hypotheses, aims and literature. Examiners were instructed on the quality levels applied to the global rating scale, which was in common usage as a performance rating scale throughout the medical course.

Study Design

This study utilised a quantitative, between groups design to investigate the impact of transitioning from a single generic rubric used by all examiners, to using one of six task-specific rubrics on the concordance between the marks of the two examiners assigned to mark the research project.

Procedure

Two examiners marked each research thesis over a three-week period in the middle of the year. All examiners are nominated by the research supervisors and usually assess one thesis (and no more than two) per year in one or more of the five broad categories of medicine, women's health and paediatrics, surgery and anaesthesia, community. population and global health, and other areas (such as medical education). More than 350 research projects are examined annually, with more than 600 individual examiners involved in any one year. The research projects are marked by the examiners independently without any collaboration between examining pairs. Both the generic and task-specific rubrics included written information on completing the rubric; however, beyond these instructions there was no specific training provided to examiners on examining the research thesis. Examiners received all documents electronically and were responsible for completing the rubric and submitting their marks within the specified time frame. Where there was no discordance between examiners (that is, the two assessments were within 10 marks of each other), students' final mark was calculated as the average of the two marks. An adjudication could be initiated when there was discordance between examiner marks. In circumstances where the discordance between examiners was greater than 20 marks (out of 100), a third examiner was automatically assigned to remark the research project to provide an adjudication. In selected circumstances where the discordance was less than 20 marks, an adjudicator was assigned at the discretion of the subject coordinator. For adjudicated projects, the student's final mark was calculated as either the average of the two marks, or the adjudicator could select one or the other examiner's grade, using the examiner's comments to guide their decision. In this situation, the adjudicator was required to provide a brief justification of why their decision was the most appropriate resolution.

For two years, all examiners assessed the research project using the generic rubric. In the latter two years of this study, task-specific rubrics were assigned to examiners based on one of six research categories. Analyses of student achievement data to inform quality improvement of teaching and learning is conducted routinely as part of quality assurance processes for the course and do not require formal institutional ethics approval. Such analyses utilize completely anonymous data, are conducted by researchers independent of the course coordinator, and are reported at the group level only.

Data Analysis

We focussed exclusively in this study on the available data (total scores from each examiner) gathered to inform decision-making for the research thesis. Thus, the lack of rubric component scores or coding of task-specific rubric type limited the opportunity to examine the role of these characteristics in examiner scoring variation.

The analysis examined differences between examining pairs in the original percentage final marks submitted for each research thesis for the four years of this study. A one-way analysis of variance was used to explore variation in the absolute value of the difference in scores between examining pairs with planned contrasts to compare the average discrepancy between and within rubric types. A measure of inter-rater

reliability was calculated as the intra-class correlation using one-way random effects between examining pairs for the generic and specific rubrics. Further chi-square analyses explored whether the frequency of adjudication and the size of adjudication discrepancies were related to use of the generic or a task-specific rubric.

Results

Over the four years of the study, 1272 research projects were examined. Table 1 shows the averages for the absolute value of the discrepancy in percentage marks awarded by examiners for the first two years of the study when a generic rubric was used, compared with the two subsequent years when the task-specific rubrics were used. A one-way ANOVA showed significant variation in average examiner marking discrepancies across the four years, $F(3, 1271) = 19.12$, $p < 0.001$. Planned contrasts showed the average marking discrepancy was not significantly different for the two years the generic rubric was used ($M = 6.26$ in the first year and $M = 6.79$ in the second year, $p = 0.322$); however, the average mark discrepancy for the first year the task-specific rubrics were used ($M = 9.83$) was significantly larger than in the following year ($M = 8.60$, $p = 0.021$). The average mark discrepancies for both years the task-specific rubrics were used were significantly larger than for both years the generic rubric was used. Intra-class correlations as a measure of inter-rater reliability for the generic rubric was 0.38 and 0.41 for the first and second year, respectively. In contrast, the inter-rater reliability for the task-specific rubrics was lower; 0.27 and 0.36 for the first and second year, respectively (Table 1). Such values are generally regarded as poor,[20] but are comparable to the values for similar research projects reported in the literature.[18]

Overall, the number of adjudications across four years was 252 (or 19.73% of all research projects examined). Table 2 shows that the percentage of adjudications required was lower for the generic rubric ($n = 32$, 10.22% for the first year and $n = 27$, 8.36% for the second year) compared with the task-specific rubrics ($n = 113$, 35.53% for the first year and $n = 80$, 24.77% for the second year). A chi-square analysis showed the number of adjudications undertaken was not independent of the type of rubric, $\chi^2(3) = 99.55$, $p < 0.001$. Examination of the adjusted standardized residuals showed there were significantly fewer adjudications than expected when the generic rubric was used and significantly more than expected when the task-specific rubrics were used.

The total number of adjudications undertaken where the discrepancy between the original examiners was more than 20 marks ('large' discrepancies) was 69 (or 5.43% of all research projects examined). The proportion of large discrepancies was also related to the type of rubric, $\chi^2(6) = 51.62$, $p < 0.001$. Large discrepancies were less common when the generic rubric was used ($n = 7$, 2.25% for the first year and $n = 3$, 0.95% for the second year) and more common when the task-specific rubric was used, particularly in the first year of implementation ($n = 35$, 11.01% for the first year and $n = 24$, 7.43% for the second year). The balance of evidence suggests greater consistency between examiners when the generic rubric was used.

Discussion

Our study provides valuable empirical evidence of the impact of using generic compared with task-specific rubrics on the consistency of examiner judgements for the assessment of a research thesis in a medical course. Research theses are often a core form of written assessment task in medical training, but there is a notable lack of research on examiner judgements for research assessments.[14] Moreover, the need to generate empirical data to guide rubric design and implementation has been noted.[19] In developing task-specific rubrics, we prioritized examiner requests for more detailed and specific guidance on examining specific types of research project. Although there was surprisingly little evidence to guide the decision, we hypothesized more specific rubrics would be better targeted to the project type, could improve the consistency of examiner judgements and the rigour of assessment, and thus reduce the burden of escalating to a third marker in the event of marking discrepancies.

Our study demonstrated, contrary to examiner belief, that the concordance between examiner judgements for the research project was higher when the generic rubric was used for assessing the research thesis. Inconsistency between examiners in the average discrepancy between examiners (to almost 10% of the total available marks) in the first year of implementation of the task-specific rubrics was particularly notable. Although unfamiliarity with the new rubrics may underlie some initial increase in variability between examiners, the greater than expected discordance between examiners persisted in the second year of implementation. Generic rubrics were only used for two years prior to the introduction of task-specific rubrics and no specialized training accompanied either rubric; thus, the findings are not attributable to greater familiarity with the generic rubric.

Variability of written assessment marks at tertiary institutions has received considerable attention, and it has been demonstrated that experienced and knowledgeable examiners can assign different marks to the same piece of work.[21–25] Substantial effort has been invested in improving fairness, consistency and transparency in written assessment by using tools that explicitly describe assessment criteria. Rubrics are believed to assist students and supervisors arrive at a common understanding of expectations and requirements, and achieve more objective and consistent assessment of student writing.[9, 26] Yet there is a lack of evidence that rubric type promotes greater consistency of examiner judgements[16] and there is almost no evidence focussed on comparing rubric types for research projects. Our study, however, suggested that

Table 1. Descriptive Statistics for the Absolute Value of the Discrepancy between Examiner Pairs for the Generic and Task-specific Rubrics.

		Mean	SD	Median	Min	Max	N	ICC
Generic rubric	**Year 1**	6.26	5.47	5.00	0	32	312	0.38
	Year 2	6.83	4.91	6.00	0	27	317	0.41
Specific rubric	**Year 1**	9.83	8.32	8.00	0	42	318	0.27
	Year 2	8.60	7.47	6.00	0	39	323	0.36
Total		7.89	6.84	6.00	0	42	1270	0.36

ICC, Intra class correlation.

transferring from a generic to task-specific rubrics resulted in significantly decreased concordance between examining pairs for the research thesis. The particularly high discordance for the task-specific rubrics in the first year of implementation (compared with the previous two years using the generic rubric) could be explained by examiners being unfamiliar with the new rubrics. Certainly, the implementation of the new task-specific rubrics was not supported by any specialized training beyond written instructions. However, the average discordance between examining pairs remained higher than the generic rubric in the second year of implementation of the task-specific rubrics, despite greater familiarity with these rubrics.

Our approach to peer review for the research thesis accords with academic peer review for submission to journals. Assessment of the submission by two independent markers against a rubric in these scenarios is designed to reduce bias. Instead, the discordance between examiners was greater for the task-specific rubrics and thus the number of arbitrations that involved the need for a third marker also increased. It may be that generic rubrics allow tacit but shared high-level expectations of academic quality to be applied in evaluating the research thesis, whereas the greater subject and/or methodology-related detail in the task-specific rubrics could encourage examiners to apply their own methodological practices and expectations to the marking. A similar idea is included in alternative descriptions of generic rubrics as 'top-down'[14] or 'universal'[4], suggesting such rubrics tap into common understandings about quality in research that may defy detailed explication. This is one of the central debates in rubric design and implementation[19] and, while intriguing, this explanation of the above finding remains speculative.

Examiner marking behaviour is also only one facet of the function of rubrics. Rubrics are also designed to provide guidance to students in interpreting their performance.[27] The educational implications of different rubric types and student engagement with rubrics for learning and development purposes is an important aspect of future research.[14,28,29]

Table 2. Number and percentage of research project adjudications undertaken for the generic and the task-specific rubrics.

		Total Adjudications		>20 Adjudications	
		N	%	N	%
Generic rubric	**Year 1**	32	10.22	7	2.25
	Year 2	27	8.36	3	0.95
Specific rubric	**Year 1**	113	35.53	35	11.01
	Year 2	80	24.77	24	7.43
Total		252	19.73	69	5.44

Limitations

The study provides valuable real-world empirical evidence on the relationship between rubric type and examiner marking consistency for a research project in a medical course. Nonetheless, there are several limitations associated with the research. One interpretation of the increased discordance in examiner marks is the use of task-specific rubrics; however, other interpretations are plausible. For instance, two years' experience marking the research thesis using the generic rubric may have established a mental model among examiners of performance standards for the thesis. Thus, on initial implementation of the task-specific rubrics, some examiners may have continued to apply their previous understanding of the generic rubric and adapted their marking using the task-specific rubrics to fit. Significant variation in how the task-specific rubrics were used may be likely in the absence of explicit in-person examiner training in how to use the task-specific rubrics, despite familiarity with the global rating scale and written instructions on its use.

Overall, the rubrics differed fundamentally along generic versus task-specific lines; nonetheless, the rubrics also differed systematically in several areas that may have contributed to the further discordance for the task-specific rubrics. For instance, the use of fixed ranges for performance categories for the generic rubric (which may have encouraged a more holistic approach) compared with scoring of individual items to generate a category score for the task-specific rubrics. The generic rubric design also employed H1-H2A-H2B-H3-N

categories, compared to Excellent-Good-Satisfactory-Borderline-Fail for the task-specific design. For the generic rubric design, descriptive terms were used to designate each level of performance within categories. In contrast, the specific rubrics relied on the examiner's judgment as to how well they considered each domain was answered. This may explain the higher degree of variability in this set of rubric data. That is, examiners different interpretations of what an Excellent-Good-Satisfactory-Borderline-Fail looked like could have contributed to greater mark variance.

Although we suggest overall that task-specific rubrics were associated with greater variability in the marks awarded by examining pairs, it is possible that certain types of task-specific rubric used in this context had greater discordance between examiners than others. Unfortunately, across the period in which these marks were tracked there was no specific coding in the data files to capture the type of task-specific rubric used to mark the research thesis. Further analysis to determine whether different task-specific rubrics are responsible for the increased discordance is warranted. These analyses should also consider examining the properties of individual scales within the rubrics to explore variation in examiner marks for different scale types and for different components of the research thesis.

Conclusions

Our study contributes significantly to the evidence base on the impact of rubric type on the consistency of examiner judgements for a research project in a medical education context. The unexpected finding in the current study that generic rubrics promoted greater consistency between examiners than task-specific-rubrics may highlight an important feature of generic rubrics, namely an underlying tacit understanding among examiners of what broadly constitutes quality in research. Perhaps in aiming to be more explicit, we broke down this tacit understanding only to introduce greater variation at the level of individual methodological practice. However, rubrics are complex structures and in redesigning the rubrics for our research projects, we also introduced other structural changes that may have influenced examiner concordance. For better or worse, educational practice and reform is often opportunistic and can involve multiple simultaneous changes which, while pragmatic, does not create ideal conditions for comparative research. We believe our study points to a potentially significant finding about rubrics in the academic research context, but one which will need further investigation and ideally replication in other, preferably better controlled contexts.

REFERENCES

1. Reddy YM, Andrade H. A review of rubric use in higher education. *Assess Eval High Edu*. 2010;35(4):435-448. doi:10.1080/02602930902862859
2. Popham WJ. What's right - and what's wrong - with rubrics. *Educ Leadership*. 1997;55(2):72-75.
3. Hack C. Analytical rubrics in higher education: a repository of empirical data. *Brit J Edu Technol*. 2015;46(5):924-927. doi:10.1111/bjet.12304
4. Timmerman BEC, Strickland DC, Johnson RL, Payne JR. Development of a 'universal' rubric for assessing undergraduates' scientific reasoning skills using scientific writing. *Assess Eval High Edu*. 2011;36(5):509-547. doi:10.1080/02602930903540991
5. Moskal BM, Leydens JA. Scoring rubric development: validity and reliability. *PARE*. 2000;7(10):1-6. doi:10.7275/q7rm-gg74
6. Grierson L, Winemaker S, Taniguchi A, Howard M, Marshall D, Zazulak J. The reliability characteristics of the REFLECT rubric for assessing reflective capacity through expressive writing assignments: a replication study. *PME*. 2020;9(5):281-285. doi:10.1007/s40037-020-00611-2
7. Hansson EE, Svensson PJ, Strandberg EL, Troein M, Beckman A. Inter-rater reliability and agreement of rubrics for assessment of scientific writing. *Education*. 2014;4(1):12-17. doi:10.5923/j.edu.20140401.03
8. Hayward MF, Curran V, Curtis B, Schulz H, Murphy S. Reliability of the interprofessional collaborator assessment rubric (ICAR) in multi source feedback (MSF) with post-graduate medical residents. *BMC Med Educ*. 2014;14(1049):1-9. doi:10.1186/s12909-014-0279-9
9. Jonsson A, Svingby G. The use of scoring rubrics: reliability, validity and educational consequences. *Educ Res Rev*. 2007;2(2):130-144. doi:10.1016/j.edurev.2007.05.002
10. Cook D, Brydges R, Ginsburg S, Hatala R. A contemporary approach to validity arguments: a practical guide to kane's framework. *Med Educ*. 2015;49(6):560-575. doi:10.1111/medu.12678
11. Kane MT. Validation. In: Brennan RL, ed. *Educational Measurement*. 4th ed. American Council on Education and Praeger Publishers; 2006: 17-64.
12. Schuwirth LW, Van der Vleuten CP. Programmatic assessment: from assessment of learning to assessment for learning. *Med Teach*. 2011;33(6):478-485. doi:10.3109/0142159X.2011.565828
13. Dawson P. Assessment rubrics: towards clearer and more replicable design, research and practice. *Assess Eval High Edu*. 2017;42(3):347-369. doi:10.1080/02602938.2015.1111294
14. Prins FJ, Kleijn RD, Tartwijk JV. Students' use of a rubric for research theses. *Assess Eval High Edu*. 2017;42(1):128-150. doi:10.1080/02602938.2015.1085954
15. Cyr P, Smith K, Broyles I, Holt C. Developing, evaluating and validating a scoring rubric for written case reports. *IJME*. 2014;5:18-23. doi:10.5116/ijme.52c6.d7ef
16. Brookhart SM. Appropriate criteria: key to effective rubrics. *Front Educ*. 2018;3(22):8–19. doi:10.3389/feduc.2018.00022
17. Chong A, Romkey L. Testing inter-rater reliability in rubrics for large scale undergraduate independent project. Procedings of the 2016 Canadian Engineering Education Association (CEEA16) Conference. 2016; Paper 105. Accessed March 28, 2021. https://ojs.library.queensu.ca/index.php/PCEEA/article/view/6465
18. Williams L, Kemp S. Independent markers of master's theses show low levels of agreement. *Assess Eval High Edu*. 2019;44(5):764-771. doi:10.1080/02602938.2018.1535052
19. Panadero E, Jonsson A. A critical review of the arguments against the use of rubrics. *Educ Res Rev*. 2020;30(100329):1-19. doi:10.1016/j.edurev.2020.100329
20. Cicchetti DV. Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. *Psych Assess*. 1994;6(4):284-290. doi:10.1037/1040-3590.6.4.284
21. Bloxham S, den-Outer B, Hudson J, Price M. Let's stop the pretence of consistent marking: exploring the multiple limitations of assessment criteria. *Assess Eval High Edu*. 2016;41(3):466-481. doi:10.1080/02602938.2015.1024607
22. Hand L, Clewes D. Marking the difference: an investigation of the criteria used for assessing undergraduate dissertations in a business school. *Assess Eval High Edu*. 2000;25(1):5-21. doi:10.1080/713611416
23. Oakleaf M. Using rubrics to assess information literacy: an examination of methodology and interrater reliability. *J Am Soc Inf Sci Tech*. 2009;60(5):969-983. doi:10.1002/asi.21030
24. Saunders MN, Davis SM. The use of assessment criteria to ensure consistency of marking: some implications for good practice. *Qual Assur Educ*. 1998;6(3):162-171. doi:10.1108/09684889810220465
25. Stellmack MA, Konheim-Kalkstein YL, Manor JE, Massey AR, Schmitz JAP. An assessment of reliability and validity of a rubric for grading APA-style introductions. *Teach Psych*. 2009;36(2):102-107. doi:10.1080/00986280902739776
26. Ahmed A, Pollitt A. Improving marking quality through a taxonomy of mark schemes. *Assess EducX*. 2011;18(3):259-278. doi:10.1080/0969594X.2010.546775
27. Gearhart M, Herman JL, Novak JR, Wolf SA. Toward the instructional utility of large-scale writing assessment: validation of a new narrative rubric. *Assess Writ*. 1995;2(2):207-242. doi:10.1016/1075-2935(95)90013-6
28. Beeth M, Cross L, Pearl C, Pirro J, Yagnesak K, Kennedy J. A continuum for assessing science process knowledge in grades K-6. *EJRSME*. 2001;5(3). https://ejrsme.icrsme.com/article/view/7657
29. Luft J. Rubrics: design and use in science teacher education. *J Sci Teach Educ*. 1999;10(2):107-121. doi:10.1023/A:1009471931127

Appendix

Appendix 1: Generic rubric

Category	H1 (80-100)	H2A (75-79)	H2B (70-74)	H3 (65-69)	N (<65)	Maximum score	Allocated score
Abstract	Clearly and concisely written. Contains key findings, major methods and results. Significance of study and conclusions presented clearly.	Clearly and concisely written. Contains most key findings, methods and results. Significance of study and some conclusions presented.	Clearly written summary. Contains findings, methods and results. Significance of study and some conclusions included.	Contains key findings, major methods and results. Conclusions and significance of study not presented clearly.	Contains key findings, but with poor description of methods and results. Conclusions and significance of study not described.	10	
Introduction	Highly focussed and concise background, leading to explanation of context and perspective. Clear links between hypotheses, aims, purpose and literature.	Focussed and concise background, leading to a clear overview but lacks perspective. Links between hypotheses, aims, purpose and literature.	Background is not focussed or concise, lacks completeness. Links between aims, purpose and literature, but no clear hypotheses.	Much of the key basic information missing in background. No clear links between hypotheses, aims, purpose and literature.	Little or no critical review of the articles cited. No discussion of the strengths and weaknesses of the highlighted studies. No hypothesis and/or aim provided.	15	
Methodology	Clear and detailed description of methods and statistical analysis. Statistical analysis is appropriate, accurate and presented clearly.	Clear description of methods and statistical analysis. Statistical analysis is appropriate and accurate but minor detail lacking.	Description of methods and statistical analysis mostly clear but significant detail lacking. Statistical analysis is appropriate but minor inconsistencies.	Description of methods and statistical analysis lacking major details. Statistical analysis is limited and has major inconsistencies.	Poor description of methods. Little or inappropriate statistical methods used.	10	
Results*	Arranged logically with data presented clearly in text, tables and figures with standalone legends. No labelling errors.	Data presented clearly in text, tables and figures with standalone legends.	Data presented in figures and text. Descriptive figure and table legends.	Data presented in figures, tables and text. Errors in labelling and poor figure and table presentation.	Negligible or excessively tedious reporting of results. Long lists of tables and graphs (if any) which serve little purpose.	20	
Discussion	Logical and comprehensive discussion. Clear understanding of the significance of the data. All major themes included. Critical approach and supporting evidence discussed.	Discussion clear and logical. Most major themes included. Evidence of a critical approach and understanding of the significance of the data.	Narrative style without critical approach. Few links between data and published work. Some major omissions in discussion.	Discussion does not extend beyond results, Misunderstanding of some major concepts. Limited critical analysis of experiments and no clear links.	Major gaps in key material. The student's understanding of the area is marginally adequate and often inaccurate.	25	
Conclusions	Summarises key arguments and provides vision for the future.	Conclusions supported by data. Includes summary but lacks vision of future.	Lacks completeness, attempts summary. Misalignment between conclusions and data. Few future directions identified.	Major misalignment between conclusions and data. Few future directions identified. No vision or link back to aims.	No conclusion provided at all or so poor as to be worthless.	10	

(continued)

Continued.

Category	H1 (80-100)	H2A (75-79)	H2B (70-74)	H3 (65-69)	N (<65)	Maximum score	Allocated score
Organisation and presentation	References cited correctly in text with correct formatting in reference list. Attractive layout with subheadings and illustrations to emphasize ideas. Negligible typographical and grammatical errors. Appropriate discipline- specific terminology, abbreviations and writing style. Word limit is within guidelines	References cited correctly in text with consistent but incorrect formatting in reference list. Use of appropriate font and layout with illustrations to emphasize ideas. Some subheadings. Few typographical and grammatical errors. Generally appropriate discipline-specific terminology, abbreviations and writing style. Word limit is within guidelines.	Some errors in citing and formatting reference list. Acceptable font and layout without illustrations and subheadings. Some typographical and grammatical errors. Occasional misuse of discipline-specific terminology, abbreviations and writing style. Word limit is within guidelines.	References not cited correctly in text. Errors and inconsistent formatting of reference list. Inappropriate font and layout without illustrations. Typographical and grammatical errors. Consistent mis-use of discipline-specific terminology, abbreviations and writing style. Word limit is outside guidelines.	Difficult to read. Important topics omitted and badly organized. References missing and significant errors in formatting of reference list, Multiple spelling and/or grammatical errors which affect understanding. Poor layout. Frequent mis-use of discipline-specific terminology, abbreviations and writing style. Word limit is significantly outside guidelines.	10	
Total:						**100**	

Appendix 2: Task-specific rubric example

Research Project Assessment Sheet – Meta-analysis		Excellent 5	Good 4	Satisfactory 3	Borderline 2	Fail 1	SCORE
Abstract	Is the abstract a structured summary?						
	Is it clearly and concisely written?						
	Does it contain the overall protocol?						
	Are the significance and conclusions of the study clearly presented?						
Introduction	Is the introduction highly focussed?						
	Does it contain a concise background leading to an explanation of the research question?						
	Are the hypotheses clearly stated?						
	Are the aims clearly stated?						
	Are there clear links between the hypotheses, aims and literature?						
Methodology	Is there a clear description of the methods?						
	Have the key words for the search strategy and names of online search databases been clearly described?						
	Has the specific inclusion and exclusion criteria been described?						
	Has the process of data abstraction and quantitative data synthesis (principal measure of effect, methods of combining results, handling of missing data and assessment of statistical heterogeneity) been clearly described?						

(continued)

Continued.

Research Project Assessment Sheet – Meta-analysis		Excellent 5	Good 4		
Satisfactory 3	Borderline 2	Fail 1		SCORE	
	Has the assessment of publication bias been clearly outlined?				
Results	Has a PRISMA diagram been clearly presented?				
	Were the included studies listed along with important characteristics and results of each study?				
	Were the findings of the individual studies combined appropriately?				
	Have the results of the systematic review been reported in an orderly manner and included important information on the applicability of evidence?				
	Have results for each treatment group in each trial, for each primary outcome, and data needed to calculate effects on size and confidence intervals been presented?				
	Has the risk of bias within included studies been reported?				
	Have stand-alone legends (containing no errors) been used?				
Discussion	Have the main findings been summarized including the strength of evidence for each main outcome?				
	Is there a critical discussion of the results?				
	Have limitations of the study, outcome level and review been described?				
Conclusion	Have the key findings/arguments been presented?				
	Has the significance of the study been stated?				
Organisation and Presentation	Have all of the relevant subheadings suggested by the PRISMA guidelines been included?				
	Have the references been cited correctly in the text?				
	Has the reference list been formatted correctly?				
	Is the layout attractive, with clear subheadings and illustrations to emphasise ideas?				
	Are there any typographical and grammatical errors?				
	Has appropriate, discipline-specific language been used?				
	Has the correct formatting been used?				