ARTICLE

Applying a Rubric Development Cycle for Assessment in Higher Education: An Evidence-based Case Study of a Science Communication Module

Brenda Pui Lam YUEN and Sirinut SAWATDEENARUNAT

Centre for English Language Communication, National University of Singapore

Correspondence:

Name: Brenda YUEN
Address: Centre for English Language Communication, National University of Singapore, 10 Architecture Drive, Singapore 117511
Email: elcbypl@nus.edu.sg

Recommended Citation:

Yuen, B. P. L., & Sawatdeenarunat, S. (2020). Applying a rubric development cycle for assessment in higher education: An evidence-based case study of a science communication module. *Asian Journal of the Scholarship of Teaching and Learning, 10*(1). 53-68.
https://doi.org/10.24112/ajsotl.103127

ABSTRACT

Although empirical studies on the use of rubrics have been undertaken in a wide range of disciplines and for various purposes in higher education, surprisingly little attention has been paid to the process of rubric development to establish the quality of rubrics, especially in the assessment of science communication. In the context of a new science undergraduate module where this study is undertaken, it is essential to develop and validate the rubric that can assess the extent to which students are able to attain the learning outcomes to ensure that students' learning outcomes are accurately and consistently assessed. This study presents a validation study of a task-specific assessment rubric developed for a communication module for science undergraduates using an embedded mixed-method design. The results of the Many-facet Rasch Model analysis indicate that the rubric appears to be functioning well, with the raters, items and rating scale functioning as intended by the model; however, the scale descriptors need fine-tuning. The qualitative analysis of tutors' feedback also supports revision and recommends refinement of the descriptors. The results from the surveys show that students found the rubric useful in helping them understand their achievement levels and enhance their writing performance. This study not only provides direction for future revisions of the rubric, but also confirms the importance of a robust process of developing and validating rubrics in higher education using a rubric development cycle, especially when a large cohort of students and multiple raters are involved.

Keywords: Rasch measurement model; rubric design; rubric validation; science communication

RUBRIC FOR POPULAR SCIENCE NEWS GENRE

In the science communication module *Exploring Science Communication through Popular Science* revamped by the Centre for English Language Communication (CELC) at the National University of Singapore, a task-specific rubric was constructed based on a 4-step approach to rubric construction which involves: i) choosing the assessment method; ii) identifying *evaluative criteria*; iii) defining *scoring strategies*; and iv) describing *quality definitions*. Those four steps involve the three key rubric features, *evaluative criteria*, *scoring strategies*, and *quality definitions* as identified by Popham (1997) and Dawson (2017). Evaluative criteria are the factors that a rater considers when determining the quality of student's work. Scoring strategies involve the use of a rating scale to interpret judgments of student's performance. Quality definitions provide a detailed explanation of the skills that a student should demonstrate in order to attain a particular level of achievement.

The popular science news genre, as a type of written genre in the science communication field, has been chosen as the assessment task. Students are expected to re-contextualise scientific ideas for a non-specialist audience, and deploy appropriate textual and linguistic features/strategies in popular science communication. The task requires students to engage and inform non-specialist readers of a new development, innovation or breakthrough in science.

When identifying evaluative criteria, the identification of writing skills specific to popular science news genre was a challenge because of a lack of empirical research about the assessment of popular science news writing in higher education. Despite numerous empirical studies on rubric use in a wide range of disciplines and for various purposes in higher education (Reddy & Andrade, 2010), no literature has been found on rubric design and validation for science communication. Therefore, concepts in popular discourse (Calsamiglia & van Dijk, 2004; Myers, 1991) and media discourse (Bednarek, 2006) were explored. Bell's (1991) *news values* in media discourse has been influential in linguistics (Bednarek, 2006; Durant & Lambrou 2009; Bednarek & Caple, 2012). Bell (1991) defines *news values* as "the often unconscious criteria by which news workers make their professional judgements as they process stories" (p.155) and categorises *news values* as *brevity*, *clarity* and *colour*. These three key values are commonly known as 'news writing objectives' (Bednarek & Caple, 2012, p. 40). As they are related to the quality or style of the news text, *clarity* and *colour* have been adopted as the two major evaluative criteria. The first evaluative criterion, *Clarity*, was sub-categorised into *Accessibility* and *Organization of ideas,* whereas the second criterion, *Colour*, was sub-divided into *Significance of the key finding,* and *Language strategies to appeal and engage readers.*

An analytic rating scale has been identified as the scoring strategy. This type of scale has been proven to reduce rater variability by limiting assessment criteria to specific constructs (Hamp-Lyons, 1991, 1995; Knoch, 2009; Ahmed & Pollitt, 2011). As for the optimal number of performance categories, Miller (1956) argues that the number of performance categories raters can differentiate is between seven plus or minus two levels. Research studies (North, 2000, 2003; Myford, 2002) have concurred that reliability was highest for scales ranging from five to nine scale points. Therefore, an analytic 8-point rating scale with 5 qualitative categories was used in this study. Scores were differentiated into 5 categories, namely *Inadequate* (Band 1), *Developing* (Band 2), *Satisfactory* (Bands 3 and 4), *Competent* (Bands 5 and 6), and *Very Good* (Bands 7 and 8).

Quality definitions were carefully crafted based on critical analysis of authentic popular science news articles published in renowned popular science magazines. Quality definitions for each of the evaluative criteria are shown in Table 1.

Table 1
Quality definitions for key rubric items

	Criteria	Quality definitions
Clarity	Accessibility (25%)	• Explanation of the key finding: the scientific concepts and processes are accessible to non-specialist readers (through the use of explanatory strategies and clarifying techniques) • Selection of materials: materials are relevant and sufficient for non-specialist readers to understand, but not too technical • Grammar/syntax: errors occur only as 'slips' and do not hinder meaning
	Organization of ideas (25%)	• Position of the moves: moves are strategically placed to enhance understanding • Macro-organization: moves are logically presented to establish a clear link between sections • Micro-organization: no logical gaps exist between explanations and conclusions
Colour	Significance of the key finding (25%)	• The implication of the key finding to science and the public: the implication is stated and explained. • The rationale of the study: The rationale is clearly stated (the background to the problem is clearly stated along with present solutions and their limitations) • The significance of the key finding: the significance is appropriately appraised through the use of evaluative language and booster language
	Language strategies to appeal and engage readers (25%)	• Use of appeals to entice readers • Writing style/tone: the style is appropriate to the popular science news genre (non-academic register) • Use of evaluative language (unexpectedness) to engage readers

OBJECTIVES

The purpose of this study is to validate a task-specific rubric in order to provide implications for rubric development in the context of science communication in higher education. The following three research questions were formulated:

i) How psychometrically valid and reliable is the rubric?
ii) What are raters' perceptions of the efficacy of the rubric on grading?
iii) What are students' perceptions of the use of the rubric on improving performance?

METHODS

Participants

The study comprised 334 science undergraduates aged between 18 and 22 who enrolled in the science communication module, and nine experienced course instructors as raters. Out of 334 students, 150 completed the survey, 98 (65.3%) of whom indicated English as their first language. All respondents had reviewed the rubric before completing the task, and used it for self-assessment and peer-assessment.

Instruments

Performance ratings. Performance ratings were graded using the rubric. Each article was marked by a single rater on four criteria on an eight-point scale. All raters marked an additional seven scripts to ensure linkage of the data set so that the severity measures constructed for the raters could be compared to one another (Eckes, 2008, 2011).

The sample scripts. A set of five sample scripts was provided to raters. Scripts were double-marked and selected according to three aspects: i) the topic of source texts (*Life Science* versus *Statistics*), ii) performance levels with special emphasis across two qualitative performance categories (e.g. *Satisfactory* and *Competent*) and sub-bands within one category (e.g. Bands 5 and 6–*Competent;* Bands 7 and 8–*Very Good*); and iii) evaluative criteria to probe and address raters' understanding of specific descriptors. Raters were asked to justify their ratings in writing.

The interview questions. General questions focused on the raters' use of the rubric, while specific questions were based on their actual scores of the five sample scripts. These questions were primarily to investigate their perception of the evaluative criteria and descriptors.

The survey. To measure students' perceptions of rubric use, the survey included 16 Likert-scale items and 2 open-ended questions. These questions elicited students' responses of their perceived clarity and usefulness of the rubric and its effectiveness in enhancing their writing performance. Eight statements were designed to assess the four aspects of the usefulness of the rubric, namely *anxiety*, *self-efficacy*, *self-regulation*, and *transparency*. These eight statements were weighted on a 6-point Likert scale ranging from "Completely disagree" (1) to "Completely agree" (6). *Anxiety* was measured by statements reflecting a reduction in levels of anxiety when the rubric was applied ("The rubric helped lower my anxiety when doing these assignments"; "The use of the rubric made me feel more confident when I worked on these assignments"). *Self-efficacy* was assessed with statements targeting students' capability to achieve a specific goal, for instance, their ability to identify one's strengths and weaknesses or reflect on one's work ("The rubric helped identify my strengths and weaknesses when writing these assignments"; "The rubric helped me reflect on my work before submitting these assignments"). To measure *self-regulation*, participants were asked whether the rubric supported their self-regulation by planning and monitoring their progress ("The rubric was helpful in planning my approach to these assignments."; "The rubric helped me check my work in progress."). *Transparency* was measured by statements about increased transparency by establishing expectations and requirements ("The rubric helped me understand what was expected from me for these assignments"; "The rubric clarified the components and requirements of these assignments").

Students' responses were averaged out to compute the indices of the overall usefulness and the four aspects of usefulness. Another eight items were used to identify students' perceived clarity and usefulness of the descriptors on understanding their level of performance in terms of the four evaluative criteria. The first four items regarding the clarity of the descriptors were weighted on a 5-point Likert scale ranging from "Not clear at all" (1) to "Extremely clear" (5), while the second four items regarding the usefulness of the descriptors were weighted on a 5-point scale ranging from "Not useful at all" (1) to "Extremely useful" (5).

Procedures

This study is situated in the paradigm of embedded mixed-methods research in which qualitative data collected would be used to supplement quantitative data to support the findings. The team collected three types of data. First, performance ratings were analysed using the Many-facet Rasch Model (MFRM) by FACETS 3.71.3 (Linacre, 2013). Second, nine raters participated in a two-hour face-to-face training session to ensure their understanding of the rubric. 334 science news articles were marked using the rubric, of which 63 were double-marked to achieve the necessary connectedness in the data. Once the rating process concluded, five raters agreed to rate and provide commentaries for the sample scripts, and were interviewed individually. Each semi-structured interview lasted one and a half to two hours, and was audio-recorded and transcribed for data coding and analysis. Third, students' surveys on their perceptions of the rubric were administered online at the end of the semester.

RESULTS

Analysis of students' ratings

The Many-facet Rasch Model (MFRM) refers to a class of measurement models that extends the basic Rasch model (Rasch, 1980) by incorporating more variables or facets than the two that are typically included in a testing situation, i.e. examinees and items (Eckes, 2009). In this study, a three-facet Rasch model was used. The three facets in the analysis were: student ability (334 elements); rater severity (9 elements); and the difficulty of the items (3 elements).

MFRM analysis provides estimates of measures for the calibration of students' ability, raters' severity, items' difficulty, and the quality of the rating scale. Measures were expressed in logits. The variable map shown in Figure 1 presents the variable map, a graphic representation of the spread of student measures (ability), rater measures (severity), item (difficulty), and the location of the thresholds for the rating scale categories, on the same logit scale. In the first column, the measurement spans from -12 to 10 logits. The second column shows the distribution of students. Each "*" represents 5 students, and each "." represents 1 to 4 students. The majority of students scored above "0", indicating their task performance was satisfactory. The third column indicates raters 2, 3, 4, 6, 7 and 8 are more severe, whereas raters 1, 5 and 9 are more lenient. The fourth column shows that *Accessibility* was the most difficult item, while *Language strategies* was the easiest. The last column shows the levels of the scale as a whole.

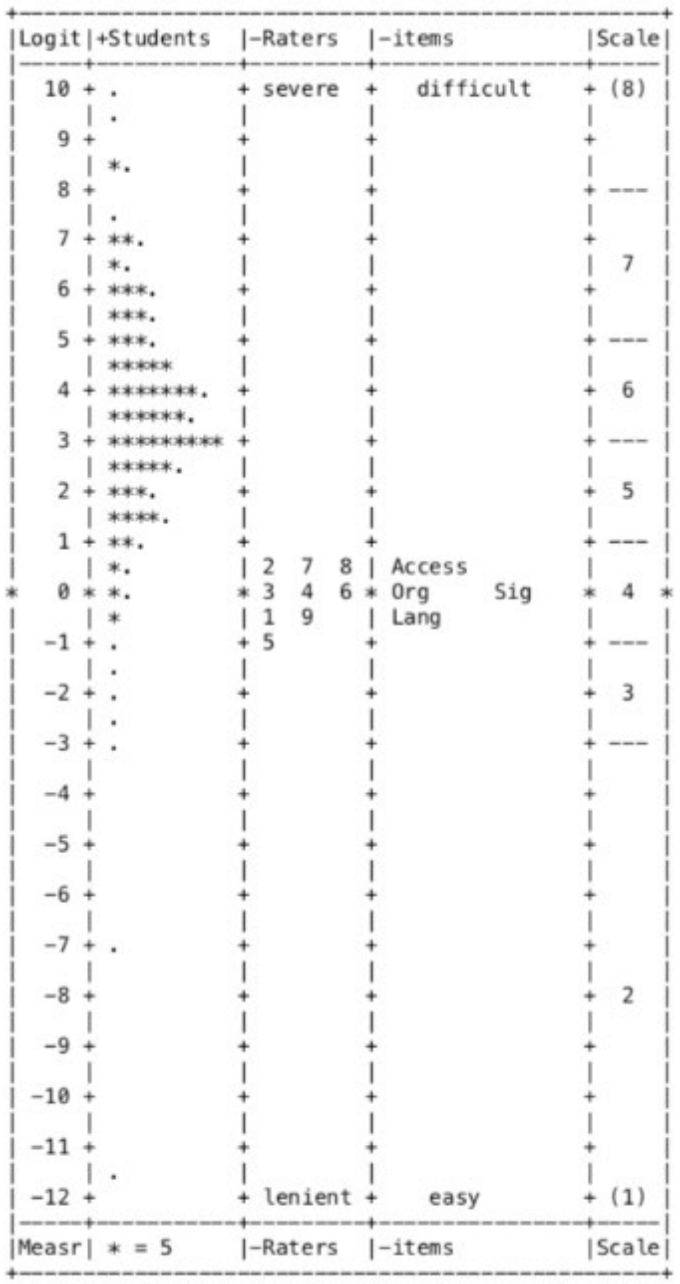

Figure 1. Variable map.

Table 2 shows the rank ordering of raters based on levels of severity and presents fit statistics to identify the degree to which raters used the scale of the rubric in a consistent manner. The raters varied in their measures of severity, ranging from -1.21 to +0.71 logits. Rater 7 was the most severe (logit measure = 0.71) and Rater 5 was the most lenient (logit measure = -1.21). Raters with fit values greater than 1 show more variation than expected in their ratings, whereas raters with fit values less than 1 show less variation than expected (Eckes, 2009). All raters, except Rater 5, had mean-square fit statistics that stayed within Linacre's (2008) defined acceptable range of 0.50 and 1.50. Although Rater 5 showed a heightened degree of misfit with a mean-square Infit value of 1.63, the other 8 raters performed in a consistent manner.

Table 2
Measurement results for the rater facet

Raters	Severity Measure	S.E.	Infit	Outfit	Observed Average	Fair Average	Number of Ratings
7	0.71	0.12	1.24	1.22	5.93	5.42	168
2	0.58	0.09	0.96	0.94	5.28	5.48	296
8	0.47	0.11	0.86	0.88	5.46	5.54	168
6	0.25	0.12	0.88	0.86	5.69	5.65	160
3	0.16	0.1	0.7	0.7	5.61	5.68	228
4	-0.21	0.15	1.11	1.11	5.68	5.86	100
1	-0.26	0.12	0.98	1.01	5.82	5.88	152
9	-0.49	0.12	0.86	0.85	5.93	5.98	160
5	-1.21	0.12	1.63	1.62	5.89	6.28	156

Note: S.E.is Standard error. Infit and outfit are mean-square statistics.

Table 3 presents the rank ordering of items based on levels of difficulty. The items are ranked from the most difficult at the top to the least difficult at the bottom. The fair average ratings range from 5.44 for *Accessibility* to 6.03 for *Language strategies*. In other words, *Accessibility* is the most difficult item and *Language strategies* is the least difficult item. All four items are functioning as intended by the model for items as mean Infit and Outfit mean-square statistics are near the expected value of 1.

Table 3
Measurement results for the item facet

Items	Difficulty Measure	S.E.	Infit	Outfit	Observed Average	Fair Average	Number of Ratings
Accessibility	0.67	0.07	0.89	0.90	5.36	5.44	397
Organization of ideas	0.05	0.07	1.08	1.07	5.64	5.73	397
Significance of the key finding	-0.13	0.08	0.90	0.90	5.72	5.82	397
Language strategies to engage readers	-0.60	0.08	0.08	1.11	5.92	6.03	397

Note: S.E. is Standard error. Infit and outfit are mean-square statistics.

Engelhard and Wind's (2013) guidelines for examining the rating scale quality were applied to verify the quality of each rating scale category. The seven guidelines include: i) directionality, ii) monotonicity, iii) category usage, iv) distribution of ratings, v) rating scale fit, vi) category coefficient order, and vii) category coefficient locations. When the guidelines are met, the rating scale categories can be used to describe student location on a latent variable (Linacre, 1999).

In terms of the first guideline *Directionality*, as shown in Table 4, the categories appear to be aligned with the latent variable due to the close match between observed and expected average measures for all categories. *Directionality*, also known as the coherence of category usage, is defined as a match between observed and expected ratings (Linacre, 2002). In terms of the second guideline *Monotonicity*, a monotonic progression of rating scale categories is indicated by the agreement between the observed and expected ordering of the categories. The increasing rating scale categories correspond to the increasing average person measures on the latent variable. In terms of the third and fourth guidelines *Category Usage* and *Rating Distributions*, frequency and distribution of observed average measures show a good spread of ratings across all categories except for Category 1. While Guideline 3 focuses on frequencies of observations within categories, Guideline 4 refers to the percentage of observations. The lowest category, *Inadequate*, has only one observation, which is below the minimum of 10 observations needed for each category as suggested by

Linacre (2002). Despite the infrequent observation for Category 1, the ratings for other categories generally conform to a normal distribution. In terms of the fifth guideline *Rating Scale Fit*, the requirement is met as outfit mean-square statistics across all categories were below 2. High outfit mean-square value of more than 2 would mean excessive randomness, suggesting the use of categories in unexpected contexts (Engelhard & Wind, 2013). A reasonably uniform level of randomness was observed for Categories 2 to 8 with their outfit mean-square values being close to an expected value of 1; however, Category 1 has a relatively low outfit mean-square value of 0.3. In terms of the sixth guideline *Category Coefficient Order*, a monotonic increase of category coefficient order was observed, with the lowest category coefficient location being identified between Categories 1 and 2 (-13.2 logits), and the highest location between Categories 7 and 8 (7.97 logits). In terms of the seventh guideline *Category Coefficient Locations*, the category distinction guideline had been violated between Categories 2 and 3 as the difference between the 1/2 category coefficient location (-13.2 logits) and 2/3 category coefficient location (-2.71 logits) has an absolute value of 10.49 logits, which is above the maximum difference of 5 logits as provided by Linacre (2002).

Table 4
Category statistics for the rating scale

Rating Scale Category	Label	Category Usage (%)	Average Measure		Outfit	Thresholds		
			Observed	Expected		Measure	S.E.	Difference
1	Inadequate	1 (0%)	-12.71	-12.11	0.3			
2	Developing	19 (1%)	-5.03	-4.79	0.6	-13.20	1.17	
3	Satisfactory (low)	46 (3%)	-0.63	-0.61	0.9	-2.71	0.35	10.49
4	Satisfactory (high)	186 (12%)	1.15	1.10	1.1	-1.14	0.18	1.57
5	Competent (low)	395 (25%)	2.48	2.53	0.9	1.08	0.10	2.22
6	Competent (high)	558 (35%)	3.92	3.85	1.0	2.84	0.07	1.76
7	Very good (low)	322 (20%)	5.33	5.43	1.1	5.16	0.08	2.32
8	Very good (high)	61 (4%)	7.25	7.18	0.9	7.97	0.16	2.81

Note: Outfit is a mean-square index. Thresholds are Rasch-Andrich thresholds. S.E. is standard error.

Analysis of raters' commentaries and interviews

Based on the analysis of raters' commentaries and semi-structured interviews, four major issues related to raters' perceptions of the efficacy of the rubric on grading were identified: i) the difficulty of awarding either band within a qualitative performance category; ii) the conflict of overlapping of *Accessibility* and *Organization of ideas*; iii) the clarity of the descriptors for *Accessibility*; and iv) the potential misrepresentation of students' ability for *Accessibility* and *Organization of ideas*.

The first issue is associated with the option of awarding either of the two bands within a qualitative performance category. Raters identified the correct qualitative performance category for most of the annotated scripts, but they tended to exhibit their own preference on how to award either band within one qualitative category. One interviewee pointed out that raters were given 'leeway' related to the way in which they awarded a higher or lower band.

> "As I said, we tend to have quite a broad quite a lot of leeway. Sometimes it becomes subjective. It's just like I choose one here and maybe I choose one here, but the person doesn't go here; stays somewhere in this range." (Interviewee 5)

Some raters found the differentiation between the two bands difficult, especially with the two qualitative categories, *Competent* (Band 5 or 6) and *Very Good* (Band 7 or 8). An interviewee explained their difficulty:

> "The splitting of 5 to 6 for me is more difficult between the split between 7 and 8. I mean for my own script, for example, there was a script that I gave a 5 under *Significance* because what to me was lack of appeal, whereas, for another, it was more the academic tone that should not have been there. So if we were to divorce those, I'm not sure how that might pan out; the lack of flexibility there." (Interviewee 4)

Another salient point made by the interviewees as a challenge when using the rubric is the potential conflict of the overlapping of *Accessibility* and *Organization of ideas* due to the lack of certain rhetorical moves. While A*ccessibility* involves the information presented in rhetorical moves, *Organization* is associated with the arrangement of those moves. Interviewees indicated that the absence of some moves could affect the performance of both *Accessibility* and *Organization of ideas*, leading to the problem of double-penalisation.

> "For me, the tricky thing is with *Accessibility* and *Organization*. The 'missing moves' part, you know. So, is it the content or is it the logical flow and the gap? That's there. So, I think when we talked about it we said okay, if you penalize the student in one area, we don't double penalize. That's the thing we keep in mind currently." (Interviewee 2)

> "If there could be clarity between the moves and the logical gaps, why we put the moves under *Accessibility* and the logical gaps under *Organization*, if that could be clearer then there wouldn't be the question of penalizing the student twice." (Interviewee 5)

The third issue that has been identified is the clarity of two specific descriptors for *Accessibility* in terms of explanation of the key finding. The lack of progression from *Satisfactory* ("the scientific concepts and processes still remain *slightly technical*") to *Competent* ("the explanation of the scientific concepts or processes may be *too simplistic* or readers may need some background knowledge to understand") was identified through the use of judgemental words such as 'slightly technical' and 'too simplistic'. One interviewee illustrated this point:

> "But they (the descriptor under *Satisfactory*) still remain slightly technical. And then the next one when we go to *Competent*, talks about explanations which maybe too simplistic. If it is simplistic should it still be competent or should it be the other way around wherein the strategies are used but they remain slightly technical." (Interviewee 5)

The last issue was the potential misrepresentation of students' ability for *Accessibility* and *Organization of ideas.* To achieve a high level of *Clarity*, information in the article must be consistently accessible to non-specialist readers and presented in a logical manner. However, raters commented that some news articles were penalised one band because of a flaw in one section of the article. One interviewee addressed their concern for a fair assessment of student ability based on their overall performance:

> "Sometimes some articles are really good, and it's a pity because they missed on move 1 or 2 and then you have to really put them down. That's another scenario—like it's not fair but some of the… or it could be implicit but it's such a good article, very well written." (Interviewee 3)

Analysis of students' survey

In Table 5, the mean scores for the perceived clarity of the descriptors in the four domains, i.e. i) *Accessibility*; ii) *Organization of idea*; iii) *Significance of the key finding*; and iv) *Language strategies,* ranged from 3.85 to 3.94 on a 5-point response continuum. This implies that students found the descriptors in all domains generally clear. In terms of perceived clarity, the lowest was *Organization of ideas* (3.85) and the highest mean was the *Significance of the key finding* (3.94). In terms of understanding one's performance using the descriptors, *Accessibility* has the lowest score of 3.75 while *Significance of the key finding* has the highest at 3.88.

Correlational analyses were used to examine the relationship between clarity of the descriptors and the understanding of their performance using the descriptors. Results of the Spearman correlation in Table 5 indicated strong positive correlations between clarity and understanding of one's performance were found (r=0.74 for *Accessibility*; r=0.73 for *Organization*; r=0.77 for *Significance*; r=0.78 for *Language strategies*, p<0.01).

Table 5
Correlations between clarity and usefulness of the descriptors by domains (n=150)

	Clarity		Usefulness		Correlations
	Mean	Std. Deviation	Mean	Std. Deviation	
Accessibility	3.87	0.82	3.75	1.02	0.74**
Organization of ideas	3.85	0.82	3.78	1.00	0.73**
Significance of the key finding	3.94	0.79	3.88	0.99	0.77**
Language strategies to engage readers	3.89	0.80	3.78	0.98	0.78**

**. Correlation is significant at the 0.01 level (2-tailed).

Table 6 presents descriptive statistics and internal consistencies of the four constructs for the rubric use scale. Cronbach's alpha for the anxiety, self-efficacy, self-regulation, and transparency scales were 0.86, 0.88, 0.89, and 0.90 respectively. The mean score for transparency was the highest at 4.97 on a 6-point response continuum while the lowest was for anxiety with a score of 4.43.

Table 6
Descriptive statistics and internal consistencies of the Sub-dimensions of Usefulness scale (n=150)

	Mean	Std. Deviation	Cronbach's alpha
Anxiety	4.43	1.17	0.86
Self-efficacy	4.84	1.04	0.88
Self-regulation	4.81	1.04	0.89
Transparency	4.97	0.98	0.90

Students' responses to both open-ended questions indicate an overall positive perception of the rubric use. 80% of students stated that the use of the rubric had enhanced their writing performance. 73.7 % of students indicated that no change to the rubric was needed. To further analyse students' perceptions of the usefulness of the rubric, their responses to the first open-ended question were categorised based on the four constructs of rubric usefulness. The prominent ideas indicated in their responses coincide with the quantitative survey results. Students who indicated improvement in their writing performance, cited increasing transparency and supporting their self-regulation as the major reasons. Some representative comments are as follows:

> "The rubric tells me exactly *what is required* of me in my writing, so that I can *ensure that all the components of a good piece of writing are present* in my work."

> "The rubrics gave me a clear understanding of *what is required* of me, and *guided* me as I kept the rubrics in mind when writing and *checking* my work."

> "*Describing clearly what was expected of me* for the news articles was very useful in *helping me to plan and decide* where to put my attention-grabbers, the moves 1-8 and also my teleological appeals."

DISCUSSION

First, results of the MFRM analysis indicate that the rubric appears to be functioning well, with the raters, items and rating scale functioning as intended by the model, but the descriptors need to be refined. In terms of raters' performance using the rubric, all raters except Rater 5 displayed some degree of variability in their level of severity within the acceptable range. They were internally consistent, suggesting an individual consultation with Rater 5 is necessary to understand their interpretation of the standards and criteria. In terms of item difficulty, *Accessibility* was the most difficult for students to achieve a certain score. It is essential to identify and understand the reasons for such a phenomenon. In terms of the functioning of the rating scale, the quantitative description of rating scale categories suggests adherence to most of Engelhard and Wind's (2013) guidelines. Despite an acceptable usage and distribution of ratings across seven categories, the ratings for the lowest category, Category 1, do not meet Guidelines 3 (*Category Usage*) and 4 (*Rating Distributions*) because of its extremely low usage. Guideline 7's (*Distinct Category Coefficient Locations*) deviation is indicated by a large absolute value of the difference between category coefficient locations for Categories 2 and 3. This implies that a large gap existed between Category 2 and 3; therefore, scale descriptors can be modified to make the *Developing* band more difficult and the *Satisfactory* band easier.

Second, qualitative analysis of raters' commentaries and interviews also support revision and recommends refinement of the descriptors. Most raters identified performance categories correctly for the annotated scripts; however, they exhibited their own preference on how to award the higher or lower bands within a qualitative performance category. Raters identified some descriptors as more salient than others, and awarded bands based on their preference. In addition to the difficulty of awarding bands, raters commented that the lack of certain rhetorical moves might have double-penalised students in terms of *Accessibility* and *Organization of ideas* and suggested specific moves should be explicitly stated in the rubric for each performance category. Raters also indicated that some students were penalised by one band because of a flaw in one particular section, emphasising the lack of clarity and progression in the descriptors under *Accessibility*.

Third, the results from the surveys show that students found the rubric useful in helping them understand their achievement levels and enhancing their writing performance. In terms of rubric use, improved transparency was perceived as the most useful aspect of the rubric, while the least useful was reduction in anxiety. *Organization of ideas* had the lowest mean scores in terms of perceived clarity of the descriptors,

suggesting students may not find the descriptors for *Organization of ideas* clear enough. *Accessibility* had the lowest mean scores in terms of the perceived usefulness of the descriptors on understanding one's performance level, implying that students might not be able to understand their performance in that domain. The strong positive correlation indicates that the higher the degree of perceived clarity, the higher the level of perceived usefulness of understanding one's level of performance. A very high percentage of students stated that the use of the rubric had enhanced their writing performance and indicated no change was needed.

As indicated by raters and students, *Accessibility* and *Organization of ideas* are the two evaluative criteria that would need refining. The three descriptors in both items should be arranged according to their degree of importance and the most salient descriptor should be highlighted in the rubric to help raters decide how to award either band. *Accessibility* and *Organization of ideas* can be graded by sections to avoid any potential misrepresentation of students' ability. This will help students achieve a clearer understanding of their actual performance for different sections and reflect on their performance specifically. To tackle the issue of clarity and progression of the descriptors for *Accessibility*, non-judgemental terminology should be used consistently across performance categories. Once the issue of clarity is tackled, students will find the rubric more useful in understanding their achievement levels and would use the rubric to assess their performance more effectively.

IMPACT ON THE RUBRIC

Re-categorization of the rubric items

In response to the conflict arising from the overlapping of *Accessibility* and *Organization of idea*s, and the potential misrepresentation of students' abilities for those two rubric items, *Clarity* has been re-categorised into: i) *Context of the study*; and ii) *The reported study.* The descriptors for *Accessibility* and *Organization of ideas* have been merged and students' performance in terms of *Clarity* would be assessed by those two sections. To help students understand the scope of the two new rubric items, definitions and explanations specifying the necessary rhetorical moves were provided, as indicated in Table 7. Specifying the moves in both sections will allow students and teachers to understand the criteria and standards more effectively.

Table 7
Quality definitions for *Clarity*

Criteria		Quality definitions
Clarity	Context of the study (20%)	The context of the study refers to the background information that helps readers understand why this particular study is needed (what is the gap) or what has led the researchers to conduct the study. This context includes (but not limited) to the background of the study (i.e. Move 3) and the context and rationale leading to the objective of the study (i.e. Move 4). • The explanations of the context are greatly tailored to the assumed knowledge base of potential readers through use of explanatory strategies, suitable and sufficient information, and appropriate word choice.* • All ideas are presented coherently and logically which leads to an understanding of the objective of the study. • The writing is very fluent; the author shows a good control of language use.
	Reported study (30%)	The reported study refers to the information about the study that supports the key finding or the main claim introduced in the headline, lead or first section of the news article (i.e. Move 1). The reported study includes (but not limited to) the methods and the results (i.e. Move 6) obtained through the Methods section, and the explanation of the results (Move 7). • The explanations of the reported study are greatly tailored to the assumed knowledge base of potential readers through use of explanatory strategies, suitable and sufficient information, and appropriate word choice.* • All ideas are presented coherently, logically which support the key finding/main claim introduced in Move 1. • The writing is very fluent; the author shows a good control of language use.

Prioritising and refining key descriptors

In response to the difficulty of awarding sub-bands within a qualitative performance category and the category disparity between *Developing* and *Satisfactory*, the descriptors have been arranged according to their degree of importance. The most salient descriptor has been highlighted with an asterisk in the rubric to emphasise that it is the necessary condition for a student to achieve the higher band within a qualitative performance category. Measurable descriptive words were incorporated to describe observable performance.

IMPLICATIONS FOR RUBRIC DEVELOPMENT IN HIGHER EDUCATION

This study has not only provided direction for future revisions of the rubric, it has also confirmed the importance of having a well-planned and well-designed process to develop and validate rubrics in higher education, especially when a large cohort of students and multiple teachers are involved. Therefore, we propose the implementation of a rubric development cycle that would help rubric developers in higher education create and validate a rubric that suits the needs and requirements of a specific educational context such as “English for Specific Purposes” courses in order to facilitate fair grading processes and enhance student learning. The cycle consists of three key stages: i) a four-step process to rubric construction; ii) rubric application with tutors and students; and iii) a mixed-method rubric validation. These three stages are in a loop that can be iterated to enhance rubric quality for assessment and grading, providing effective feedback, and promoting student learning.

In the process of rubric construction, Popham's (1997) three key features of a rubric and Dawson's (2017) identification of rubric design elements were considered in the design of our rubric. Based on this study, 5 qualitative performance categories with 2 sub-bands in the top three categories could be used in a rubric. Specific descriptors should be highlighted to help raters differentiate students' performance when sub-bands are used. In terms of rubric application, both teachers and students receive face-to-face training to understand the *quality definitions,* and the standards of performance levels. Teachers are involved in the training session, in which they apply the rubric to assess sample work in order to achieve shared understanding of the criteria and standards prior to grading. Students are also trained to use the rubric to assess the quality of work through peer and self-assessment. In the validation process, a mixed-method research design is adopted. MFRM analysis can be used to examine the psychometric quality of the rubric, including the degree of rater severity and consistency, the difficulty of items and functioning of the rating scale. Qualitative feedback from teachers and students could also be collected to identify any disparity of the interpretations of the rubric between students and teachers. By adopting the rubric development cycle—from rubric design to rubric application and validation—a valid and reliable rubric would be able to measure students' learning outcomes fairly.

CONCLUSIONS

This study has provided evidence that the rubric developed, based on the rubric development cycle, can facilitate fair grading processes, provide effective feedback, and enhance student learning in a science communication module in higher education. The results of the findings have given insights into the reliability and validity of the rubric so that the concise and precise articulation of the criteria and standards, and shared understanding and agreed consensus over the interpretation of criteria and standards among teachers and students can be achieved. This will further facilitate the effective utilisation of the rubric to help students understand the expectations of the criteria and engage them more actively with their learning processes through self-assessing their work based on the rubric.

ACKNOWLEDGEMENTS

The authors gratefully acknowledge the Centre for Development of Teaching and Learning (CDTL) for granting the Teaching Enhancement Grant in Academic Year 2017/18.

ABOUT THE AUTHORS

Brenda Pui Lam YUEN is a Lecturer at the Centre for English Language Communication (CELC), where she has taught and coordinated undergraduate courses in academic writing and critical thinking. She has also been involved in the development and validation of university-wide English language proficiency and placement tests in Hong Kong and Singapore. Her teaching and research interests include second language writing, language testing and assessment, particularly rubric validation using Rasch modelling.

Sirinut SAWATDEENARUNAT is currently working as a Learning Designer in Melbourne, Australia. She worked as a Lecturer at the CELC and coordinated the science communication module "Exploring Science Communication through Popular Science" from 2015 to 2019. Her teaching and research interests include curriculum design, assessment and learning experience.

REFERENCES

Ahmed, A. & Pollitt, A. (2011). Improving marking quality through a taxonomy of mark schemes. *Assessment in Education: Principles, Policy & Practice*, *18*(3), 259-278. https://doi.org/10.1080/0969594X.2010.546775

Bednarek, M. (2006). *Evaluation in media discourse: Analysis of a newspaper corpus*. London/New York: Continuum.

Bednarek, M. & Caple, H. (2012). *News discourse*. London, New York: Bloomsbury.

Bell, A. (1991). *The language of news media.* Cambridge, MA: Blackwell.

Brownell, S. E., Price, J. V., & Steinman, L. (2013). Science communication to the general public: why we need to teach undergraduate and graduate students this skill as part of their formal scientific training. *Journal of Undergraduate Neuroscience Education*, *12*(1). Retrieved from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3852879/.

Calsamiglia, H., & Van Dijk, T. A. (2004). Popularization discourse and knowledge about the genome. *Discourse & Society*, *15*(4), 369-389. https://doi.org/10.1177%2F0957926504043705

Dawson, P. (2017). Assessment rubrics: towards clearer and more replicable design, research and practice. *Assessment & Evaluation in Higher Education*. *42*(3), 345-360. https://doi.org/10.1080/02602938.2015.1111294

Durant, A. & Lambrou, M. (2009). *Language and media: a resource book for students*. Routledge.

Eckes, T. (2009). Many-facet Rasch measurement. In Takala, S (Ed.), *Reference supplement to the manual for relating language examinations to the Common European Framework of Reference for Languages: Learning, teaching, assessment (Section H)*. Strasbourg, France: Council of Europe/Language Policy Division.

Engelhard, G. J. & Wind, S. A. (2013). Rating quality studies using rasch measurement theory. *CollegeBoard Research Report*, *2013*(3). The College Board. Retrieved from https://files.eric.ed.gov/fulltext/ED558109.pdf.

Hamp-Lyons, L. (1991). Scoring procedures for ESL contexts. In Hamp Lyons, L. (Ed), *Assessing second language writing in academic contexts* (pp. 241-76). Norwood, NJ: Ablex.

Hamp-Lyons, L. (1995). Rating non-native writing: the trouble with holistic scoring. *TESOL Quarterly, 29*, 759–62.

Knoch, U. (2009). Diagnostic assessment of writing: A comparison of two rating scales. *Language Testing*, *26*(2), 275-304. https://doi.org/10.1177%2F0265532208101008

Linacre, J. M. (2002). Optimizing rating scale category effectiveness. *Journal of Applied Measurement*, *3*, 86-106. Retrieved from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.424.2811&rep=rep1&type=pdf.

Linacre, J. M. (2013). *Facets: A computer program for Many-facet Rasch Measurement*. Chicago: Winsteps.com.

Miller, G. A. (1956). The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological Review, 63*(2), 81-67. https://doi.org/10.1037/h0043158

Myers, G. (1991). Lexical cohesion and specialized knowledge in science and popular science texts. *Discourse processes*, *14*(1), 1-26. https://doi.org/10.1080/01638539109544772

Myford, C. M. (2002). Investigating design features of descriptive graphic rating scales. *Applied Measurement in Education*, *15*(2), 187-215. https://doi.org/10.1207/S15324818AME1502_04

North, B. (2000). *The development of a common framework scale of language proficiency*. New York: Peter Lang.

North, B. (2003). *Scales for rating language performance: Descriptive models, formulation styles, and presentation formats.* TOEFL Monograph, 24.

Popham, W. J. (1997). What's wrong - and what's right - with rubrics. *Educational Leadership*, *55*(2), 72-75. Retrieved from http://www.ascd.org/publications/educational-leadership/oct97/vol55/num02/What's-Wrong%E2%80%94and-What's-Right%E2%80%94with-Rubrics.aspx.

Rasch, G. (1980). *Probabilistic models for some intelligence and attainment tests.* Chicago: University of Chicago Press.

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