JOURNAL OF MULTIDISCIPLINARY RESEARCH AND DEVELOPMENT | ISSN 3031-9390
Vol. 2, No. 2, 2025, 215-228
https://doi.org/10.56916/jmrd.v2i2.1235

Assessing the Validity and Reliability of a Rubric-Based Assessment

Jen Hua Ling*
Centre for Research of Innovation & Sustainable Development, School of Engineering and Technology, University of Technology Sarawak, 96000 Sibu, Sarawak, Malaysia

***Corresponding Author:** lingjenhua@uts.edu.my

Keywords
Rubric
Validity
Reliability
Subjective assessment
Survey

Article History
Received 2025-04-21
Accepted 2025-07-13

Abstract
This paper presents a self-evaluation study of the Final Year Project II (FYPII) report rubric, examining its validity, reliability, and effectiveness in supporting student learning. While rubrics are widely used in education, concerns remain about their fairness and consistency in subjective assessment. This study employed a validity checklist, an explicitness check, an interrater survey, and an awareness survey to evaluate the rubric's performance. Although it met all validity criteria, ambiguous terms affected reliability. The interrater survey revealed discrepancies among raters, leading to the proposal of four rating rules to enhance consistency. Despite facilitating learning, the rubric showed a gap between students' understanding of the criteria and their ability to produce quality work. Effective supervisor supervision was identified as crucial in bridging this gap. These findings highlight that a well-designed rubric alone is insufficient—proper implementation is essential to ensure meaningful assessment without overburdening evaluators.

INTRODUCTION

Engineering students are trained to solve problems to prepare for real-world challenges. Traditionally, these problems are well-defined, structured and constrained (McNeill, Douglas, Koro-Ljungberg, Therriault, & Krause, 2016; Schraw, Dunkle, & Bendixen, 1995; Shin, Jonassen, & McGee, 2003). However, modern workplace problems are often complex and ill-structured (Jonassen, Strobel, & Lee, 2006). These open-ended challenges require analytical skills, critical thinking, and creativity. As a result, assessing students' performance in solving them is inherently subjective.

Subjective assessment raises concerns about credibility and fairness. It lacks objectivity and depends on assessors' judgment, which can vary. Different assessors may emphasize different aspects, leading to inconsistency and bias. When grades are based on overall impressions, the process lacks transparency, leaving students uncertain about the evaluation criteria. Without clear guidance, their learning is not effectively driven (Ling, 2024).

Rubrics are commonly used in subjective assessment as scoring tools for qualitative ratings of authentic or complex student work (Jonsson & Svingby, 2007). They consist of two key components: criteria and performance level descriptions (Brookhart, 2018). Criteria define important aspects of assessment (Jonsson & Svingby, 2007), while performance level descriptions describe varying degrees

of quality, from excellent to poor (Andrade, 2000). Rubrics set clear expectations, guide student learning, and promote consistent, fair grading.

The effectiveness of rubrics is debated. While many educators believe rubrics improve assessment quality (Jonsson & Svingby, 2007), they often overlook reliability issues (Rezaei & Lovorn, 2010). Rater bias can persist despite their use (İlhan, 2019), and rubrics do not always ensure valid performance judgments (Bryant, Maarouf, Burcham, & Greer, 2016). Critics argue that rubrics are overly reductive (Kavanagh & Luxton-Reilly, 2016; Rezaei & Lovorn, 2010). To improve reliability, educators may narrow assessment formats and limit what rubrics measure (Bennett, 2016), sacrificing validity (Wiggins, 1994). As a result, rubrics may fail to capture students' full potential. If an achievement falls outside the predefined criteria, it is often overlooked or disregarded (Bennett, 2016). This limits students' efforts to only what is explicitly outlined in the criteria.

Rubrics are intended to ensure consistent and fair grading while guiding student learning. However, their effectiveness depends on proper design and use, which can be challenging. Developing rubrics is often difficult (Silvestri & Oescher, 2006), requiring educators to create assessments that are both valid and reliable. They must also apply rubrics consistently during grading, which takes practice and experience. Experienced educators understand their role as assessors, use rubrics correctly, and rate students based on the stated criteria (Jeong, 2015).

A poorly used rubric can be more harmful than not using one at all (Rezaei & Lovorn, 2010). Ensuring a rubric is well-designed and properly applied is crucial for effective assessment. Despite their widespread use, rubrics face reliability and validity challenges, raising concerns about their ability to fairly assess students' performance. This paper presents a self-evaluation study on the practice of subjective assessment using a rubric, aiming to identify its limitations and suggest improvements.

METHODS

This study conducted a self-evaluation of the Final Year Project (FYP) course, offered in the final year of a four-year engineering program in Malaysia. The course spanned two consecutive semesters.

FYP assessments included reports, oral presentations, poster presentations, and supervisor evaluations (Table 1). This study focused on the FYPII report, which contributed the most to the total course marks (40%) and was assessed by at least three assessors. A moderation process ensured reliability by limiting mark variation among assessors to one-quarter of the total marks. If this threshold was exceeded, a fourth assessor was involved, and the final mark was averaged. Since it involved the most assessors compared to other assessments, it was ideal for analyzing inter-rater discrepancies and assessment consistency.

Table 1. Assessment of Final Year Project

	Assessment	Supervisor	Examiners 1 and 2	Judges 1 and 2	Total
Final year project I	Proposal Report	10	5 + 5		35
	Oral Presentation		5 + 5		
	Supervisor Evaluation	5			
Final year project II	FYPII Report*	20	10 + 10		65
	Oral Presentation		5 + 5		
	Poster Presentation			5 + 5	
	Supervisor Evaluation	5			
					100

*The focus of this self-evaluation study

Table 2. Rubric for Final Year Project II report

Criteria	CO	0 - Poor	1 - Fair	2 - Average	3 - Good	4 - Excellent	Weight
(a) Problem, Objective, Literature review	CO1, PO1	The problems are made up without awareness of contemporary issues. The rationale for carrying out the research is not stated. Serious misalignment in terms of the problems, objectives, and research design. The literature review is poorly conducted.	Identify the problems with minimal awareness of contemporary issues. The rationale for carrying out the research is not stated or unclear. Misalignment in terms of the problems, objectives, and research design. The literature review is unsatisfactory and lacks useful information for the research.	State the problems related to contemporary issues. The rationale for carrying out the research is stated but not persuasive. Slight misalignment in terms of the problems, objectives, and research design. The literature review is acceptable and somewhat useful to the research.	Point out the problems related to contemporary issues and sustainable development goals. Demonstrate sufficient engineering knowledge. The rationale for carrying out the research is clearly explained. The significance of the study is persuasive but not realistic. The problems, objectives, and research design is aligned in a logical manner. Systematically and comprehensively review literature related to the research without bias.	Critically point out the root of complex problems related to contemporary issues and sustainable development goals, which is well supported by diverse engineering knowledge. The rationale for carrying out the research is well justified by the significance of the study. The objectives are well-defined, effectively address the problems, and are coherently aligned with the research design. Critically, thoroughly, systematically, and comprehensively review the literature related to the research.	4
(b) Execute an engineering research	CO2, PO10	The research program is inappropriate. The objectives are not covered. The scope of work and research process needs to be reconstructed.	The research program is unclear or inappropriate. More than one objective is not satisfactorily covered. The scope of work and research process has many flaws, which require significant improvement.	The research program is explained but clarification is required. One objective is not satisfactorily covered. The scope of work and research process is explained but require improvement.	The research program is clearly explained. All the objectives are satisfactorily attained. The scope of work is acceptable. The research process is clearly explained.	The research program is comprehensively explained with adequate technical details. The scope of work is justified and the objectives are effectively attained. In-depth grasping of the entire research process and rigorous work are evident	3
(c) Analyse the research data	CO3, PO2	Data are obviously illogical. The analysis process is not demonstrated. The results are questionable.	Data are incomplete or not clearly presented. The analysis process is not fully demonstrated. The results may be questionable.	Data are presented but require clarification. The analysis process is unclear, incomplete, or lacking depth. The results are not evaluated for reliability.	Data is clearly presented. The analysis is completely carried out using the appropriate approach. The results are evaluated for reliability and are well documented.	Data are reliable and clearly presented. Appropriate and in-depth analysis is completely carried out using a justified approach. The results are systematically evaluated, validated or verified, and well documented	5
(d) Results and findings	CO4, PO12	The results presented are questionable. No discussion on the results. No findings are given.	The results are presented without in-depth discussion. Having difficulty identifying the findings. The research gaps are not filled.	Attempt to interpret and discuss the results but lacking proficiency. The findings are highlighted without being supported by existing	The results are correctly interpreted and discussed. The findings are supported by existing knowledge and theories. Demonstrate the ability to evaluate the impact of engineering	The results are effectively interpreted and critically and comprehensively discussed in depth. The findings are novel and significant and integrate well with existing knowledge and theories. Objectively evaluate the	6

				knowledge and theories. Attempt to fill the research gaps.	solutions. The research gaps are filled.	sustainability and impact of engineering solutions. The research gaps are effectively filled.	
(e) Conclusion	CO5, PO5	The conclusion is improperly constructed. Objectives are not addressed. The limitations of the research and recommendation for future development are not given.	The conclusion gives little insight into the research. More than one objectives are not satisfactorily addressed. The limitations of the research or recommendations for future development are not given.	The conclusion summarizes the research findings. One objective is not satisfactorily addressed. The limitations of the research and recommendations for future development are irrelevant, unpractical or not properly defined.	The conclusion clearly summarizes the research findings. Objectives are addressed. The limitations of the research are listed. Realistic recommendations for future development are proposed.	The conclusion clearly and concisely summarises the research findings as a contribution to the field. All objectives are effectively addressed. The limitations of the research are critically pointed out. Creative, realistic, and practical recommendations are proposed for future development, taking into consideration societal, health, safety, legal, and/or cultural aspects.	3
(f) Produce an academic writing	CO6, PO7	The research report is poorly written. Hardly understandable. Inappropriate terminologies are used. Poor English. The format does not meet the requirement. Inappropriate use of figures and tables.	The research report is not well written. Low readability. Inappropriate terminologies are used. Poor English and many grammatical errors. The format does not meet the requirement. A lot of errors are found in figures and tables.	The research report is written with lots of redundancy and repetitions. Low readability. Many inappropriate and imprecise terminologies are observed. Many grammatical errors. Inconsistent format, some are not following the requirement. Appropriate use of figures and tables, but some errors are found in the figures and tables.	The research report is clearly and concisely written with minimal redundancy and repetition. Readable. Points and arguments are understandable. Appropriate and precise terminologies mostly (70%-90% of the time). Minimal grammatical errors. The format is following the requirement mostly (70%-90% of the time). Appropriate use of figures and tables.	The research report is clearly and concisely written without redundancy and repetition. Highly readable and interesting to read. Points and arguments are effectively delivered. Appropriate and precise terminologies throughout (>90% of the time). Proper English and grammar are used. The format is following the requirement throughout (>90% of the time). Appropriate use of figures and tables.	4

WP1: Criteria (a); WP3: Criteria (c) + (d) + (e); WP4: Criteria (a), WP7: Criteria (a) + (b) + (c); SDG: Criteria (a) + (e)

Table 2 shows the rubric for FYPII reports, which included six assessment criteria. Each criterion had five performance levels, ranging from 0 (poor) to 4 (excellent), with distinct descriptions. The criteria had different weights, with criterion (d) carrying the highest weight (six times), followed by criterion (c) (five times), and so on. The score for each criterion was calculated by multiplying the assessor's rating by its weight. The total rubric score was 100% before being converted to 40% of the FYPII report's final grade.

The rubric was evaluated for validity, reliability, and learning facilitation (Table 3). A checklist assessed its validity (Table 4). The explicitness check examined the clarity of performance level descriptions. Each criterion was divided into sub-criteria, which were evaluated for clarity. The checklist and explicitness check ensured rubric quality.

Table 3. Scope of self-assessment process

Evaluation scope	Evaluation performed
Validity	Validity checklist
Reliability	Explicitness check and Interrater survey
Learning facilities	Awareness survey

Table 4. Validity checklist for rubric

No.	Checklist	Justifications
1.	Each criterion directly addresses a course outcome (CO) and indirectly attains a programme outcome (PO).	Constructive alignment with the CO and PO
2.	The presence of the complex engineering problem (WP) is indicated.	WP is required when it is clearly stated in the PO.
3.	The presence of sustainable development goals (SDG) is indicated.	SDG is required when it is clearly stated in the PO.
4.	All chapters in FYPII are covered.	Constructive alignment with the course content
5.	The rubric comprises three to six criteria.	Manageable numbers of criteria (Liew, Puteh, & Hamzah, 2020)
6.	The criteria are not overlapped.	To eliminate redundancies.
7.	The rating scale starts with zero (0 - Poor).	Marks are given only when qualified.
8.	The weight of each criterion is indicated.	Reflect on the emphasis of the assessment.
9.	Each performance level has a distinct description.	For discrimination against one another.

An interrater survey was conducted to analyze FYP assessors' rating behavior in different scenarios. Nine assessors participated. Assuming the rubric was explicit, seven cases were outlined (Table 5). Case 1 was clear-cut, Cases 2–4 were ambiguous, and Cases 5–7 resembled explicit but poorly designed rubrics. Respondents assigned scores to each case to simulate the rubric's interrater reliability.

The awareness survey involved 26 FYP students from the February 2023 semester batch. They completed a five-question questionnaire after submitting their FYPII reports. The survey assessed their awareness of rubrics and whether the rubric supported their learning.

The validity checklist, explicitness check, interrater survey, and awareness survey served as controls in this study. The anticipated assessment quality was compared with actual assessment results (Figure 1). The results of 26 FYP students were analyzed to determine the agreement between anticipated and actual assessment quality.

Table 5. Interrater reliability survey

Possible case	Performance criteria	Performance levels			Descriptions
		1 - Poor	2 - Moderate	3 - Good	
Case 1	Sub-criteria 1	√	√	X	All sub-criteria in levels 1 and 2 are fulfilled. None in level 3 is met.
	Sub-criteria 2	√	√	X	
	Sub-criteria 3	√	√	X	
Case 2	Sub-criteria 1	√	√	√	All sub-criteria in levels 1 and 2 are fulfilled. Some of level 3 is met.
	Sub-criteria 2	√	√	X	
	Sub-criteria 3	√	√	X	
Case 3	Sub-criteria 1	√	√	X	A majority of level 2 sub-criteria are fulfilled. A minority of level 3 sub-criteria are met.
	Sub-criteria 2	√	√	X	
	Sub-criteria 3	√	X	√	
Case 4	Sub-criteria 1	√	√	X	A minority of level 2 sub-criteria are fulfilled. A majority of level 3 sub-criteria are met.
	Sub-criteria 2	√	X	√	
	Sub-criteria 3	√	X	√	
Case 5	Sub-criteria 1	√	X	X	1 sub-criterion is fulfilled in levels 1, 2,

	Sub-criteria 2	X	√	X	and 3. None of the levels are fully met.
	Sub-criteria 3	X	X	√	
Case 6	Sub-criteria 1	√	X	X	There is a gap at level 2. A majority of level 1 sub-criteria are fulfilled. A minority of level 3 sub-criteria are fulfilled.
	Sub-criteria 2	√	X	X	
	Sub-criteria 3	X	X	√	
Case 7	Sub-criteria 1	√	X	X	There is a gap at level 2. A majority of level 1 sub-criteria are fulfilled. A minority of level 3 sub-criteria are fulfilled.
	Sub-criteria 2	X	X	√	
	Sub-criteria 3	X	X	√	

√: sub-criterion fulfilled, X: sub-criterion not fulfilled

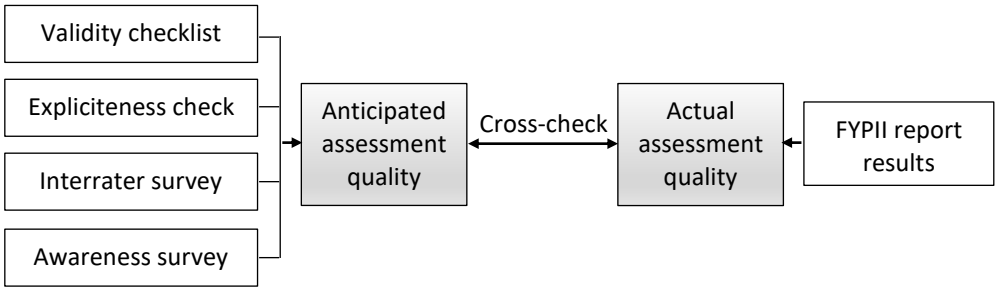

Figure 1. Evaluation of rubric-based assessment quality

RESULTS AND DISCUSSION

Validity

The rubric (Table 2) was considered valid as it met all checklist requirements (Table 4). It had six criteria, an appropriate number for assessment. Each criterion aligned with a Course Outcome (CO) and, in turn, a Programme Outcome (PO). All six COs were covered.

Among the POs, PO1 (Engineering Knowledge), PO2 (Problem Analysis), and PO5 (The Engineer and the World) required complex engineering problems (WP). The rubric implicitly addressed WP1 (Depth of Knowledge), WP3 (Depth of Analysis), WP4 (Familiarity with Issues), and WP7 (Interdependence) through various criteria. This satisfied the Engineering Accreditation Council (Engineering Accreditation Council, 2024) and International Engineering Alliance (International Engineering Alliance, 2021) requirements, which mandate WP1 and additional WPs for complex engineering problems.

PO2 and PO5 encompassed sustainable development, and thus the Sustainable Development Goals (SDGs) were therefore reflected in the rubric. The rubric covered all FYPII report chapters (introduction, literature review, methodology, results and analysis, and conclusion). Greater emphasis was placed on result analysis and research findings, so criteria (c) and (d) carried higher weights (Table 2).

The rubric applied to all FYP students, regardless of differences in supervisors, topics, and scope. While complex engineering problems (WPs) and Sustainable Development Goals (SDGs) were required, they were not directly assessed. Instead, they were embedded within the rubric rather than as stand-alone criteria. The rubric prescribed WP1, WP3, WP4, WP7, and SDGs, requiring supervisors to integrate these elements into their assigned research projects.

Reliability

During the explicitness check, each criterion was divided into two to six sub-criteria (Table 6). The descriptions were examined for clarity to ensure they distinguished performance levels from poor (rating = 0) to excellent (rating = 4). While the descriptions were generally clear, certain terms—such

as "critically," "effectively," "clearly," "comprehensively," "highly," "little," "concisely," "sufficient," "slight," "minimal," "many," and "a lot of"—were open to assessors' interpretation. These subjective terms indicated relative differences between performance levels, which could affect the rubric's reliability. To minimize ambiguity, quantifiable terms like "70%–90% of the time" could be used. Additionally, briefing FYP assessors on a standardized interpretation of rubric descriptions could further reduce rating discrepancies.

When a criterion has only one sub-criterion, interrater reliability is less of a concern. However, with two to six sub-criteria per criterion, rating discrepancies can still occur, even if the rubric is explicit and consistently interpreted. The likelihood of discrepancies increases with the number of sub-criteria. When discrepancies are unavoidable, their extent becomes critical. To assess this, an interrater survey was conducted under the assumption that the rubric was unambiguous.

Table 7 summarizes the interrater survey results, showing ratings from nine FYP assessors for various cases (Table 5). Case 1 was straightforward, with all or none of the sub-criteria met, resulting in no rating discrepancies (Diff. = 0). Cases 2, 3, and 4 were less clear-cut, involving partial fulfillment of sub-criteria at certain performance levels. These cases were common, as no rubric can always ensure clear-cut assessments. In these cases, rating discrepancies were higher (Diff. = 1.5). Cases 5, 6, and 7 typically occurred with poorly designed rubrics, where performance levels were illogically sequenced. These cases had the highest discrepancies (Diff. = 2), indicating the need for rubric revision.

Table 8 illustrates the impact of rater discrepancies on assessment outcomes. Assuming a constant discrepancy between two assessors (ranging from 0.5 to 2), the variations in final scores were calculated based on the rubric in Table 2. Each 0.5-point discrepancy resulted in a 12.5% difference in final scores, reaching up to 50% when the discrepancy was 2. This variation was primarily due to the limited number of performance levels (scale 0–4) rather than the number of criteria in the rubric. Increasing the number of performance levels could reduce final score variations.

Table 6. Explicitness check of rubric criteria

Criteria	Sub-criteria	0 - Poor	1 - Fair	2 - Average	3 - Good	4 - Excellent
(a) Problem, Objective, Literature review	(i) Problems	• made up problem • without awareness of contemporary issues	• identify problem • with minimal awareness of contemporary issues	• state probelm • related to contemporary issues	• point out problem • related to contemporary issues and sustainable development goals • sufficient engineering knowledge	• critically point out the root of complex problems • related to contemporary issues and sustainable development goals • diverse engineering knowledge
	(ii) Rationale for carrying out the research	• not stated	• not stated or unclear	• stated but not persuasive	• clearly explained • the significance of the study is persuasive but not realistic	• well justified by the significance of the study
	(iii) Problems, objective, and research design	• serious misalignment	• misalignment	• slight misalignment	• aligned in a logical manner	• objectives well-defined, • effectively address problems • coherently aligned with research design
	(iv) Literature review	• poorly conducted.	• unsatisfactory • lacks useful information	• acceptable • somewhat useful	• systematically and comprehensively • related to the research • without bias	• critically, thoroughly, systematically, and comprehensively • related to the research
(b) Execute an engineering research	(i) Research program	• inappropriate	• unclear or inappropriate	• explained • but clarification is required	• clearly explained	• comprehensively explained • with adequate technical details

	(ii) Objectives	• all not covered	• more than one not covered	• one not covered	• All attained	• effectively attained
	(iii) Scope of work	• needs to be reconstructed	• has many flaws • require significant improvement	• explained • require improvement	• acceptable	• justified
	(iv) Research work	• needs to be reconstructed	• has many flaws • require significant improvement	• explained • require improvement	• clearly explained	• in-depth grasping • rigorous work
(c) Analyse the research data	(i) Data	• obviously illogical	• incomplete or not clearly presented	• presented • but require clarification	• clearly presented	• reliable • clearly presented
	(ii) Analysis	• not demonstrated	• not fully demonstrated	• unclear, incomplete, or lacking depth	• completely carried out • using the appropriate approach	• appropriate and in-depth analysis • completely carried out • using a justified approach
	(iii) Results	• are questionable	• may be questionable	• not evaluated for reliability	• evaluated for reliability • well documented	• systematically evaluated, validated or verified, • well documented
(d) Results and findings	(i) Result interpretation	• questionable results presented • no discussion	• results are presented • without in-depth discussion	• attempt to interpret and discuss • lacking proficiency	• correctly interpreted and discussed	• effectively interpreted and critically • comprehensively discussed in depth
	(ii) Findings	• not given	• having difficulty identifying	• highlighted • without being supported by existing knowledge and theories	• supported by existing knowledge and theories. • ability to evaluate the impact of engineering solutions. • research gaps are filled	• novel and significant • integrate well with existing knowledge and theories • Objectively evaluate the sustainability and impact of engineering solutions. • research gaps are effectively filled
(e) Conclusion	(i) Conclusion	• improperly constructed	• gives little insight into the research.	• summarizes research findings	• clearly summarizes research findings.	• clearly and concisely summarises research findings • contribution to the field
	(ii) Addressing objectives	• all not addressed	• more than one not satisfactorily addressed	• One not satisfactorily addressed	• all addressed	• all effectively addressed
	(iii) Limitation	• not given	• not given	• irrelevant, unpractical or not properly defined	• listed	• critically pointed out
	(iv) Recommendation for future development	• not given	• not given	• irrelevant, unpractical or not properly defined	• proposed • realistic	• creative, realistic, and practical • consider societal, health, safety, legal, and/or cultural aspects
(f) Produce an academic writing	(i) Quality of report	• poorly written	• not well written	• written with lots of redundancy and repetitions	• clearly and concisely written • minimal redundancy and repetition	• clearly and concisely written • without redundancy and repetition
	(ii) Readability	• hardly understandable	• low readability	• low readability	• readable • understandable	• highly readable • interesting to read. • effectively delivered
	(iii) Terminologies used	• inappropriate terminologies	• inappropriate terminologies	• many inappropriate and imprecise terminologies	• appropriate and precise terminologies mostly (70%-90% of the time)	• appropriate and precise terminologies throughout (>90% of the time)

	(iv) English and grammar	• poor English.	• poor English • many grammatical errors	• many grammatical errors	• minimal grammatical errors	• proper English and grammar
	(v) Format	• not meet the requirement	• not meet the requirement	• inconsistent format • some are not following the requirement	• following the requirement mostly (70%-90% of the time)	• following the requirement throughout (>90% of the time)
	(vi) Figures and table	• inappropriate use	• a lot of errors are found	• appropriate use • some errors are found	• appropriate use	• appropriate use

Table 7. Ratings given by FYP assessors for various cases

	FYP assessors									Results*		
	A	B	C	D	E	F	G	H	I	Max	Min	Diff
Case 1	2	2	2	2	2	2	2	2	2	2	2	0
Case 2	2.5	2.5	2.5	2	2	2	2.5	2	2	2.5	2	0.5
Case 3	2	2	2	2	1.5	2	2.5	2	1	2.5	1	1.5
Case 4	2.5	2	2.5	2	1.5	2.5	2.5	2	1	2.5	1	1.5
Case 5	2	1	1.5	1	1	1.5	2	2	1	2	1	1
Case 6	1.5	1	1	1	1	1.5	1.5	1	1	1.5	1	0.5
Case 7	1.5	1	1.5	1	1	2	2.5	2	3	3	1	2

*Max = the highest rating given, Min = the lowest rating given, Diff = max – min.

Table 8. Simulated implications of rater discrepancies

Rater discrepancies	Assessor A	Assessor B	Variation (%)*[1]	Moderation process*[2]
0.5	All criteria = 2 (Score = 50%)	All criteria = 2.5 (Score = 62.5%)	12.5	Not triggered
1.0	All criteria = 2 (Score = 50%)	All criteria = 3 (Score = 75%)	25.0%	Triggered
1.5	All criteria = 2 (Score = 50%)	All criteria = 3.5 (Score = 87.5%)	37.5%	Triggered
2	All criteria = 2 (Score = 50%)	All criteria = 4 (Score = 100%)	50.0%	Triggered

*[1]Variation = score given by assessor A – score given by assessor B. [2]The moderation process was triggered when the variation exceeded 25%.

Table 7 shows that rater discrepancies can reach 1.5 even when a rubric is not poorly designed (as seen in Case 3). Table 8 indicates that moderation is triggered when discrepancies consistently exceed 1. This suggests that moderation is likely, regardless of rubric quality. A well-designed rubric alone cannot eliminate rater discrepancies.

Rater discrepancies may be reduced if the rubric is applied consistently based on agreed-upon guidelines (Table 9). These rules can make rubric assessments more objective. Table 10 presents the recommended scores for each case based on these rules. The four rules, in different combinations, effectively addressed all seven cases. However, poorly designed rubrics, such as cases 5, 6, and 7, would still require revision.

Also, the suggested rating rules have limitations. While they may improve interrater reliability, they do not necessarily ensure fair grading. This is evident in cases 5, 6, and 7, where poorly designed rubrics resulted in lower recommended scores (Table 10) compared to typical ratings in Table 7. This

raises a concern about whether students should be penalized with low scores due to flaws in rubric design.

Table 9. Rating guidelines to reduce rater discrepancies

Rules	Guideline
R1	A performance level is considered attained only when all the sub-criteria are satisfied.
R2	The quality of work in ascending order: not satisfied at all, partially satisfied, and fully satisfied. The rating should therefore adequately reflect that.
R3	When several performance levels are partially satisfied, the lowest performance level is taken.
R4	The smallest score interval is equal to half of the scale interval.

Table 10. Recommended scores based on rubric guidelines

Possible case	Performance criteria	Performance levels			Recommended score	Rating rules applied
		1 - Poor	2 - Moderate	3 - Good		
Case 1	Sub-criteria 1	√	√	X	2	R1
	Sub-criteria 2	√	√	X		
	Sub-criteria 3	√	√	X		
Case 2	Sub-criteria 1	√	√	√	2.5	R2 and R4
	Sub-criteria 2	√	√	X		
	Sub-criteria 3	√	√	X		
Case 3	Sub-criteria 1	√	√	X	2	R2 and R3
	Sub-criteria 2	√	√	X		
	Sub-criteria 3	√	X	√		
Case 4	Sub-criteria 1	√	√	X	2	R2 and R3
	Sub-criteria 2	√	X	√		
	Sub-criteria 3	√	X	√		
Case 5	Sub-criteria 1	√	X	X	1	R3
	Sub-criteria 2	X	√	X		
	Sub-criteria 3	X	X	√		
Case 6	Sub-criteria 1	√	X	X	1	R3
	Sub-criteria 2	√	X	X		
	Sub-criteria 3	X	X	√		
Case 7	Sub-criteria 1	√	X	X	1	R3
	Sub-criteria 2	X	X	√		
	Sub-criteria 3	X	X	√		

Learning facilitation

Table 11 presents the awareness survey results. All 26 students understood the purpose of a rubric and knew it was provided beforehand. Most (96.2%) reviewed it before submitting their reports, and 80.8% found it easy to understand. However, despite having the rubric, 46.2% were still uncertain about how to excel in their FYP.

Table 11. Awareness survey results on rubric-based assessment

No.	Questionnaire question	Yes	No
1.	I know what a rubric is and its function.	100%	
2.	I was aware that rubrics were provided before I submitted my FYP report.	100%	
3.	I have read through the rubrics before submitting my FYP report.	96.2%	3.8%
4.	I find the rubrics hard to understand.	19.2%	80.8%
5.	Even with the rubrics given, I am still unsure how to do well in my FYP.	53.8%	46.2%

Percentage out of 26 respondents

The questionnaire was entirely perspective-based. While student perspectives provide useful insights, they may not fully reflect reality. However, the survey results were considered reasonably reliable. Over four years, students had used rubrics in various assignments and were encouraged to review them before submission. By their final year, they were expected to understand how rubrics function. Additionally, FYP students were formally notified when the rubric was provided early in the semester.

Although students may have read the rubric, it was unclear if they had examined it thoroughly. The survey also could not rule out overconfident students who claimed to understand the rubric easily. More importantly, understanding a rubric does not necessarily translate into producing an excellent FYP report. This gap may stem partly from the complexity of FYP itself, which requires in-depth engineering knowledge, systematic research design, rigorous execution, and strong writing skills.

Therefore, in its current form, the rubric can only support student learning to a limited extent. While it sets standards and guides FYP assessors in grading, student learning still heavily relies on supervision. Without ensuring interrater reliability, the rubric risks becoming just a formal grading tool rather than an effective learning aid.

FYP results

Table 12 summarizes the individual results of 26 FYP students assessed by different evaluators. When assessors applied the rubric at their discretion, rater discrepancies ranged from 4% to 48.5%. For eight students (30.7%), discrepancies exceeded one-quarter of the total marks, triggering the moderation process and requiring a moderator (the fourth assessor) for each.

Three assessors appeared sufficient for rubric assessment. Despite notable rater discrepancies, moderation resulted in only slight changes to average scores (0.1% to 6%, Table 12). However, even these small changes could impact final grades, potentially shifting a student's grade up or down (e.g., from B to B+ or A- to B+).

The current system of three assessors is sufficient to minimize grading bias. Involving a fourth assessor increases educators' workload without significantly impacting students' final scores. To ensure moderation is only triggered, when necessary, the threshold could be reconsidered. Instead of 25% (1/4 of total marks), a threshold of 33.3% (1/3) could be more appropriate.

Using Table 12, the impact of this adjustment can be simulated. With a 33.3% threshold, the number of students requiring moderation drops from eight to five. Among them, three had pre- and post-moderation score differences exceeding 3%, likely causing a grade change. However, student no. 21, with a 26% rater discrepancy, was missed despite a 4.1% score difference. Meanwhile, students nos. 23 and 26 were flagged, though their score differences were minimal (0.4 and 0.1). Raising the threshold to 33.3% improves efficiency but may slightly compromise fairness.

Table 12. Rater discrepancies by individual student

		Grading results (%)*[1]			To determine the rater discrepancies			To gauge the effect of moderation process		
No	Supervisor	Examiner 1	Examiner 2	Moderator*[2]	Max score (%)	Min score (%)	Rater discrepancies (%)	Pre-moderation average (%)	Post-moderation average (%)*[3]	Differences
	(1)	(2)	(3)	(4)	(5) = max{(1), (2), (3)}	(6) = min{(1), (2), (3)}	(7) = (5) – (6)	(8) = average{(1), (2), (3)}	(9) = average{(1), (2), (3), (4)}	(10) = |(8) – (9)|
1	55.0	47.5	41.5		55.0	41.5	13.5	48.0		
2	58.0	58.0	52.5		58.0	52.5	5.5	56.2		

3	72.0	72.0	63.5		72.0	63.5	8.5	69.2		
4	67.5	68.5	61.0		68.5	61.0	7.5	65.7		
5	65.0	63.5	80.0		80.0	63.5	16.5	69.5		
6	64.5	41.5	58.0		64.5	41.5	23.0	54.7		
7	84.5	79.0	69.0		84.5	69.0	15.5	77.5		
8	82.0	80.5	86.0		86.0	80.5	5.5	82.8		
9	82.5	78.0	70.5		82.5	70.5	12.0	77.0		
10	45.0	43.5	54.0		54.0	43.5	10.5	47.5		
11	69.5	77.0	75.0		77.0	69.5	7.5	73.8		
12	63.5	61.0	58.5		63.5	58.5	5.0	61.0		
13	55.0	51.0	54.0		55.0	51.0	4.0	53.3		
14	72.0	65.5	56.5		72.0	56.5	15.5	64.7		
15	67.0	67.0	55.0		67.0	55.0	12.0	63.0		
16	87.5	83.5	77.0		87.5	77.0	10.5	82.7		
17	70.5	66.5	61.0		70.5	61.0	9.5	66.0		
18	69.0	74.5	77.0		77.0	69.0	8.0	73.5		
19	69.5	45.5	41.0	45.5	69.5	41.0	28.5	52.0	50.4	1.6
20	89.0	62.5	40.5	40.0	89.0	40.5	48.5	64.0	58.0	6.0
21	40.0	61.0	66.0	72.0	66.0	40.0	26.0	55.7	59.8	4.1
22	64.0	40.5	69.5	66.5	69.5	40.5	29.0	58.0	60.1	2.1
23	60.5	23.0	31.0	40.0	60.5	23.0	37.5	38.2	38.6	0.4
24	81.0	61.0	47.5	40.0	81.0	47.5	33.5	63.2	57.4	5.8
25	95.0	50.0	78.0	56.0	95.0	50.0	45.0	74.3	69.8	4.5
26	78.0	49.0	41.0	55.5	78.0	41.0	37.0	56.0	55.9	0.1

*[1]Full mark = 100%; [2]A moderator was involved when the rater discrepancies exceeded 20%; [3]Post-moderation average is computed when there was a moderation process.

Alternatively, applying the rubric guidelines (R1, R2, R3, and R4) from Table 9 could improve interrater reliability and reduce the need for moderation. Table 13 shows that after implementing these rules, cases with discrepancies exceeding 25% decreased, leading to fewer moderation cases in subsequent semesters (Sept 2023, Feb 2024, and Sept 2024). However, this finding may require further verification, as the high discrepancies in Feb 2023 could have resulted from initial unfamiliarity with the newly adopted rubric rather than the absence of guidelines, suggesting a learning curve in its application.

Table 13. Impact of rubric guidelines on moderation cases across semesters

Semester	Feb 2023	Sept 2023	Feb 2024	Sept 2024
Rubric usage	First implementation	Second	Third	Fourth
Guideline provided (R1 to R4)	X	√	√	√
Number of Students in FYP II	26	5	15	1
Number of Moderation Cases	8	2	0	0
Moderation Rate (%)	30.7%	40%	0	0

CONCLUSION

This study evaluated a rubric for Final Year Project II (FYPII) report assessment, focusing on validity, reliability, and learning facilitation. The analysis, conducted through a validity checklist,

explicitness check, interrater survey, and awareness survey, aimed to identify limitations and improve assessment quality.

Designing a rubric that is both valid and reliable while fostering learning is challenging. A well-structured rubric alone is insufficient and proper implementation is essential. For a valid assessment, FYP must align with the rubric's components (COs, POs, WPs, and SDGs) regardless of research topics. Since assessors interpret performance levels at their discretion, reliability may be affected, highlighting the need for a shared understanding among FYP assessors.

Even with a well-defined rubric, rater discrepancies cannot be entirely eliminated. A set of rating rules was introduced to improve objectivity, but their limitations must be recognized. A poorly designed rubric may unfairly penalize students. While rubrics can aid learning, understanding the criteria does not guarantee a strong FYP report; effective supervision remains crucial.

This study advises against excessive assessment when the impact is minimal, as it unnecessarily increases the workload for FYP assessors. Three assessors were generally sufficient, as the fourth assessor had little influence on final grades. Adopting rubric guidelines and raising the moderation threshold could streamline the process while maintaining reliability and fairness.

REFERENCES

Andrade, H. G. (2000). Using rubrics to promote thinking and learning. *Educational Leadership, 57*(5), pp. 13-18. Retrieved from https://ascd.org/el/articles/using-rubrics-to-promote-thinking-and-learning

Bennett, C. (2016). Assessment rubrics: Thinking inside the boxes. *Learning and Teaching, 9*(1), pp. 50-72. https://doi.org/10.3167/latiss.2016.090104

Brookhart, S. M. (2018). Appropriate criteria: Key to effective rubrics. *Frontiers in Education, 3*(22), pp. 1-12. https://doi.org/10.3389/feduc.2018.00022

Bryant, C. L., Maarouf, S., Burcham, J., & Greer, D. (2016). The examination of a teacher candidate assessment rubric: A confirmatory factor analysis. *Teaching and Teacher Education, 57*, pp. 79-96. https://doi.org/10.1016/j.tate.2016.03.012

Engineering Accreditation Council. (2024). EAC Manual Engineering Programme Accreditation Standard 2024: Board of Engineers Malaysia. Retrieved from https://eac.org.my/v2/wp-content/uploads/2024/08/Engineering-Programme-Accreditation-Standard-2024.pdf

İlhan, M. (2019). An empirical study for the statistical adjustment of rater bias. *International Journal of Assessment Tools in Education, 6*(2), pp. 193-201. https://doi.org/10.21449/ijate.533517

International Engineering Alliance. (2021). Graduate Attributes & Professional Competencies, Ver 4: 21 June 2021. Retrieved from https://www.ieagreements.org/assets/Uploads/IEA-Graduate-Attributes-and-Professional-Competencies-2021.1-Sept-2021.pdf

Jeong, H. (2015). Rubrics in the classroom: do teachers really follow them? *Language Testing in Asia, 5*(1), 6. https://doi.org/10.1186/s40468-015-0013-5

Jonassen, D., Strobel, J., & Lee, C. B. (2006). Everyday problem solving in engineering: Lessons for engineering educators. *Journal of Engineering Education, 95*(2), pp. 139-151. https://doi.org/10.1002/j.2168-9830.2006.tb00885.x

Jonsson, A., & Svingby, G. (2007). The use of scoring rubrics: Reliability, validity and educational consequences. *Educational Research Review, 2*(2), pp. 130-144. https://doi.org/10.1016/j.edurev.2007.05.002

Kavanagh, S., & Luxton-Reilly, A. (2016). Rubrics used in peer assessment. Paper presented at the Proceedings of the Australasian Computer Science Week Multiconference, Canberra, Australia. https://doi.org/10.1145/2843043.2843347

Liew, C. P., Puteh, M., & Hamzah, S. H. (2020). Comparative study of engineering design project assessment rubrics to address the Washington Accord's complexity attributes. *ASEAN Journal of Engineering Education, 4*(1), pp. 71-94. https://doi.org/10.11113/ajee2020.4n1.21

Ling, J. H. (2024). Implementation of complex engineering problem solving projects in a Malaysian engineering programme. *Indonesian Journal of Education and Social Sciences, 3*(2), pp. 133–150. https://doi.org/10.56916/ijess.v3i2.722

McNeill, N. J., Douglas, E. P., Koro-Ljungberg, M., Therriault, D. J., & Krause, I. (2016). Undergraduate students' beliefs about engineering problem solving. *Journal of Engineering Education, 105*(4), pp. 560–584. https://doi.org/10.1002/jee.20150

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

Schraw, G., Dunkle, M. E., & Bendixen, L. D. (1995). Cognitive processes in well-defined and ill-defined problem solving. *Applied Cognitive Psychology, 9*(6), pp. 523-538. https://doi.org/10.1002/acp.2350090605

Shin, N., Jonassen, D. H., & McGee, S. (2003). Predictors of well-structured and ill-structured problem solving in an astronomy simulation. *Journal of Research in Science Teaching, 40*(1), pp. 6-33. https://doi.org/10.1002/tea.10058

Silvestri, L., & Oescher, J. (2006). Using rubrics to increase the reliability of assessment in health classes. *International Electronic Journal of Health Education, 9*, pp. 25-30. Retrieved from https://www.fahefoundation.org/wp-content/uploads/2020/04/4122-13989-1-CE.pdf

Wiggins, G. (1994). The constant danger of sacrificing validity to reliability: Making writing assessment serve writers. *Assessing Writing, 1*(1), pp. 129-139. https://doi.org/10.1016/1075-2935(94)90008-6