DEVELOPMENT AND VALIDATION OF A DIFFERENTIATED PBL ASSESSMENT RUBRIC FOR INDONESIAN ELT

I Putu Andre Suhardiana

English Language Education Department
Universitas Hindu Negeri I Gusti Bagus Sugriwa Denpasar
Indonesia
putuandresuhardiana@gmail.com

Abstract

Indonesian English Language Education department is increasingly filled with diverse learners requiring an assessment strategy, but no empirically-based rubric can be found to evaluate the Project-Based Learning (PBL) outcomes in Indonesian EFL context. This study addresses this gap by creating and validating the Differentiated Assessment Rubric for PBL (DAR-PBL), which incorporates flexible weighting of criteria and allows students to choose focus of assessment, in order to leverage learner diversity. This study was conducted in four phases (rubric development, content validity assessment with three expert judges, face and construct validity using six lecturers and 20 students as participants, and inter-rater reliability testing with 20 authentic PBL work samples scored by six raters) following DeVellis's (2017) validation framework. Results Quantitative analyses included Aiken's V for content validity, mean comparisons for face and construct validity; Gwet's consensus reliability coefficient (AC1) was used to calculate the intraclass correlation coefficients (ICC) with Fleiss' Kappa. Qualitative thematic analysis of the open-ended responses provided supplementary support. Results showed excellent content validity because five of six criteria resulted in Aiken's V > 0.75 (English Language Use :V=1.00). Face validity refers to the 4.0 mark (M=4.50 lecturers, M=4.35 students), and construct validity were also robust (M=4.61lecturers / M=4.42 students). Inter-rater reliability was good (ICC=0.87, Fleiss' Kappa=0.68) and test-retest reliability excellent (mean r=0.91). The DAR-PBL represents a valid working tool that can help Indonesian ELT lecturers assess PBL in a fair manner, with implications for differentiating assessment practices across EFL contexts.

Keywords - differentiated assessment, Project-Based Learning, rubric validation, English Language Education, Indonesian higher education

Introduction

Over the previous years, ELT in higher education sector in Indonesia has experienced a shift towards pedagogical approaches that are more student-centred and better prepare graduates to work complex jobs (Abdullah, 2020). One of these methods is Project-Based Learning (PBL) that has been widely adopted in the English Language Education departments of Indonesian tertiary education, since it promotes 21st-century competencies such as critical thinking skills, collaboration and communication. According to Thornhill-Miller et al., 2023), they call these skills 'higher-order skills.' This notes that the degree to which institutions of higher education consider the 4Cs as a learning outcome is now well-established, but valid assessment of these skills remains elusive across diverse institutional contexts. Although PBL has been popular in many Indonesian ELT classrooms, its assessments have remained traditional, depending mainly on summative tests and standardized criteria that are unsuitable to characterize the type of knowledge the project tasks create (Hustarna et al., 2024). Secondly, the assessment challenge is even more complicated in Indonesia as today not all university classroom students have the same English proficiency level / very diverse academic readiness and learning methods. A systematic review by (Religioni et al., 2024), provides evidence that a one-size-fits-all approach to assessment will not serve the diverse needs of all learners and confirms the need for

flexible and responsive assessment frameworks to be adopted in order for assessments used to promote equity and meaningful learning by skilfully engaging learners so that they can show their learning outcomes.

Although a dependency on differentiated assessment is acknowledged in diverse EFL classrooms, recent works indicate that Indonesian English Language Education departments still continuously apply more uniformed and rather ineffective assessment rubrics with the same criteria and descriptors to assess all students without making adjustments for needs or proficiency level (Hasanah et al., 2022). Despite evidence that learners in Indonesian university ELT classrooms differ greatly in terms of their basis language literacy capability, readiness to learn autonomously immediately after high school and preferred style for demonstrating understanding, a one-size-fits-all, band-aid practical approach continues. Rubrics that already exist (for the most part, even those written specifically for Project-Based Learning) typically have the assumption that students will be graded against common standards and that common criteria weights are set with respect to each of those standards, in addition to common levels of performance descriptors. In a systematic review of PBL rubric studies involving Asian EFL contexts, Phuong et al. (2023) found that almost all lacked mechanisms to accommodate learner diversity in terms of flexible criteria weighting, tiered performance descriptors / pathways through tasks. It can be concluded that rubric design in Asian ELT, which they argued remains 'fundamentally undifferentiated,' represents an obstacle to equitable assessment practices. In a similar vein, (Kupchyk & Litvinchuk, 2020), in their analysis of 62 university ELT programs across Southeast Asia reflected that less than 8% included some kind of differentiation (in process, content, product, and assessment). This result is of particular concern given the widespread endorsement that differentiated instruction has received as an overarching pedagogical framework for addressing learner diversity in mainstream education (Tomlinson, 2001); Hall, T., Vue, G., Strangman, N., & Meyer, 2004). Moreover, of the extremely few rubrics that profess to include differentiation principles, there are virtually no examples that have been empirically tested for content validity, construct validity, face validity or inter-rater reliability (Adnan et al., 2019). Validation studies are needed because even if a rubric claims to be differentiated in design, only empirical scrutiny can determine whether the rubric accurately measures what it claims to measure, if users consider it clear and fair, and if different raters use it with equanimity. In the Indonesian context, the absence of validated instruments is especially problematic given that English Language Education departments must serve a student population with diverse backgrounds, including learners from various regions with different levels of previous exposure to English. There is no quantitative empirical validation study examining the psychometric properties of a differentiated assessment rubric for assessing PBL outcomes in Indonesian ELT departments to date. Such a gap deprives lecturers of evidence-based resources to mark students fairly, whilst denying curriculum developers a validated framework to inform rubric creation going forward.

This study aims to build and validate a Differentiated Assessment Rubric for Project-Based Learning (DAR-PBL) for English Language Education departments in Indonesian universities over the gap mentioned above. The rubric aims to provide both a flexible assessment tool for lecturers and several variations in differentiation, criteria can be weighted differently. Learners can choose the focus of their assessment (they can either write an argument / present them), and lecturers

are given leeway according to the learning development needs of individual learners. This study focuses on a primary research question: To what extent is the newly developed DAR-PBL appropriate in terms of content validity, construct validity, face validity, and inter-rater reliability for assessing project-based learning competencies within English Language Education departments? In doing so, the present study adds to the literature in several respects. Theoretically, this research extends differentiated instruction theory into PBL assessment an area that has seen limited empirical investigation despite widespread theoretical discourse around differentiation in general (Tomlinson, 2017; Sun & Xiao, 2024). The study methodology also presents a generalizability validation framework which can be adopted by future EFL research when creating tests in other contexts. For university academics, the validated DAR-PBL is an evidence-based tool that can make PBL assessment fairer and more reliable.

Methodology

Design

This study used a mixed-methods research design, consistent with the framework of instrument development and validation as proposed by (DeVellis, 2017) and in accordance with the Standards for Educational and Psychological Testing (Goodwin & Leech, 2003). This design consisted of 4 components: (1) rubric development phase, followed by (2) content validity phase involving the judgment of experts; (3) face-validity, construct-validity and usability phase exercises run by lecturers and students; and through a final component consisting of (4) reliability-testing utilizing inter-rater consistency analysis. In line with this purpose, the design was suitable since the study looked into a description of an existing practice as well as into developing and empirically validating a new measuring tool for differentiating assessment practices in specific conditions: Indonesian English Language Education departments' project-based learning outcome in higher education.

Participants

Two sources of participants were used for different validation purposes. First, three experts from three different universities in Indonesia were recruited through an expert panel. Each expert has earned a doctorate (Dr.) in ELT or Applied Linguistics with specialization in at least one of three domains: English language assessment, Project-Based Learning (PBL) methodology and/or differentiated instruction. Establishes content validity. A panel deciding what content is valid Second, there were six English lecturers in Universitas Hindu Negeri I Gusti Bagus Sugriwa Denpasar that had been participated in the inter-rater reliability, face validity and construct validity. All six of the lecturers fit the inclusion criteria: minimum two years' experience teaching in an English Language Education department, currently using or having taught PBL in real time in their courses. Third, to test face validity, construct validities and practical usability used 20 undergraduate students enrolled in their 3rd and 5th semester. All students must have taken at least one course in this type of course to ensure they have experience with the assessment being validated. Fourth, twenty examples of real student work (iportfolios, research posters, pedagogical videos/teaching recordings, lesson plans) were obtained from said courses of participating lecturers to be used as reliability testing materials.

Instruments

This study developed four instruments. The first was a new instrument, the Differentiated Assessment Rubric for PBL (DAR-PBL) to validate. This rubric had six components (subject content knowledge, ability to communicate in English, critical thinking skills, working

together with others, creativity and originality in the presentation of ideas, organization and presentation), four levels of achievement (emerging = 1 out of 4; developing = 2 out 4; proficient = 3 out 4; exemplary = 4 out of 4), no set weighting for each component / choice to use alternate task pathways depending on readiness level. The second was an Expert Content Validation Sheet, which had a 4-point Likert scale (1 = not relevant and 4 = very relevant) for each rubric criterion and open-ended sections to provide comments on clarity, comprehensiveness, and representativeness of the construct. The researcher adapted this sheet from Aiken (1985) Validity form. The third instrument was a Face Validity, Construct Validity and Usability Questionnaire for lecturers and students containing five 5-point Likert-scale items (1 = strongly disagree, 5 = strongly agree) measuring clarity (face validity), fairness (face validity), perceived usefulness (construct validity), ease of use and time efficiency. There were also open-ended questions for gathering qualitative feedback to help refine the rubric. The fourth was an Inter-Rater Reliability Scoring Sheet to record scores for each of the 20 student work samples across all six rubric criteria, which included a formula for calculating both weighted raw and final scores.

Data Collection

Data was accumulated through 4 stages. In stage 1 (rubric development), the researcher synthesized relevant theoretical frameworks of differentiated assessment (Tomlinson, 2017), PBL assessment (Larmer et al., 2015), and rubric design (Brookhart, 2018) to create a foundational version of the DAR-PBL. In stage 2 (content validity), the three experts were sent an email / secure online link, containing the DAR-PBL and Expert Content Validation Sheet. Then each expert scored the relevance of every criterion from the rubric and gave written comments. The validation process had to be completed within a two-week deadline for experts, with a one-week reminder. In Stage 3 (face validity, construct validity, and usability), the six lecturers and 20 students were independently examined for computing statistics on a DAR-PBL following a brief 15-minute orientation. Three student work samples (matching to the description provided by each lecturer) were randomly assigned and used in an authentic fashion, whereby lecturers used the rubric to score the performance of those sample students from which descriptions came. All lecturers and students completed the Face Validity, Construct Validity, and Usability Questionnaire (Instrument 3) right after. In stage 4 (reliability testing), scoring by six independent raters on the final refined version of DAR-PBL. The same authentic student work samples were independently scored by these six lecturers using the final revised DAR-PBL in a set of 20 for consistency. Over a period of two weeks, each lecturer scored all 20 samples. The 20 work samples were administered randomly for each rater to control order effects, and scores were recorded using the Inter-Rater Reliability Scoring Sheet (Instrument 4).

Data Analysis

The data analysis used both quantitative and qualitative methods that were aligned with each type of evidence for validity and reliability. In terms of content validity, for quantitative analysis, Aiken's V formula was utilized to calculate a content validity coefficient for each rubric criterion (Aiken, 1985) with the minimum acceptable value of $V \geq 0.75$ when evaluated by three experts. Expert qualitative comments were utilized through thematic analysis to find suggestions for word changes, criteria addition / removal / structural changes. Mean scores and standard deviations were obtained separately for lecturers and students from Part A (Face Validity) of the Face Validity, Construct Validity and

Usability Questionnaire. In general, face validity was considered acceptable with a mean score of ≥ 4.0 on the 5-point scale. Open-ended questions were thematically analysed for patterns about how clear and fair the process was. Mean scores and standard deviations from Part D of the same questionnaire (perceived usefulness items) were computed separately for lecturers and students as a construct validity measure. An average score of ≥ 4.0 was considered to provide acceptable evidence of construct validity. To assess reliability (inter-rater agreement), we calculated the Intraclass Correlation Coefficient (ICC) using the two-way random effects model for absolute agreement (ICC 2,1), as recommended by Koo & Li (2016). ICC values of below 0.50 suggested poor reliability, 0.50–0.75 moderate reliability, 0.75–0.90 good reliability and above 0.90 excellent reliability [22]. Besides, Fleiss' Kappa (for categorical agreement on scores by performance level was reported in a more conservative measurement of inter-rater agreement that went beyond absolute scores. Statistical analyses were performed using either SPSS version 26 or JASP and set at a level of $\alpha = 0.05$ for statistical significance.

Finding and Discussion

Results

Content Validity

The content validity of the DAR-PBL was tested by three experts utilizing the Expert Content Validation Sheet (Instrument 2). Each expert has a doctoral degree in ELT / Applied Linguistics with competence in English language evaluation, Project-Based Learning (PBL) methodology, or differentiated instruction. The experts judged the relevance of the six rubric criteria on a 4-point scale (1 = not relevant, 2 = somewhat relevant, 3 = extremely relevant, 4 = highly relevant). Aiken's (1985) V formula was used to generate a content validity coefficient for each criterion, where the minimum acceptable value for three experts was $V \geq 0.75$. Table 1 shows the individual expert ratings and the computed Aiken's V coefficients for each of the six criteria.

Table 1. Aiken's V Coefficients for Each Rubric Criterion

Criterion	Expert 1	Expert 2	Expert 3	Aiken's V	Interpretation
Content Knowledge	4	4	3	0.92	Valid
English Language Use	4	4	4	1.00	Valid
Critical Thinking	4	3	4	0.92	Valid
Collaboration	3	3	3	0.67	Below threshold
Creativity & Originality	3	4	3	0.83	Valid
Organization & Presentation	4	3	3	0.83	Valid

As seen in Table 1, five of the six criteria surpassed the acceptable level of Aiken's V values of 0.75. The criterion of English Language Use got the highest coefficient (V = 1.00). It means that there is a perfect agreement of all three experts that this criterion is highly significant to measure PBL outcomes in English Language Education departments. Content Knowledge and Critical Thinking were rated V = 0.92, two experts rated these qualities as highly relevant and one expert rated them as quite relevant. Both Creativity and Originality and Organization and Presentation earned V = 0.83, with all three experts ranking these qualities as either extremely relevant or highly relevant. The lowest coefficient was reported for the collaboration criterion (V = 0.67), which is below the minimum permitted level. All three experts regarded this criterion as somewhat useful (scoring 3) and not highly relevant (score 4).

The findings of the evaluation of the rubric are shown in Table 2. Experts evaluated four global statements regarding the DAR-PBL on a 4-point scale (1 = strongly disagree to 4 = strongly agree).

Table 2. Rubric Evaluation by Experts

Statement	Expert 1	Expert 2	Expert 3	Mean
The rubric covers all the major points of PBL assessment in ELT	4	3	4	3.67
The performance level descriptors (Emerging to Exemplary) are well distinguished	3	3	4	3.33
The rubric is applicable for ELT settings in Indonesia	4	4	4	4.00
The rubric covers well the ideas of differentiated instruction	4	3	4	3.67

The three experts agreed / strongly agreed that the rubric is appropriate for the Indonesian university ELT environment (mean = 4.00) as can be seen in Table 2. The lowest mean score (3.33) was for the statement that the performance level descriptors are clearly differentiated, but was still above the midpoint of the scale.

Qualitative Findings from Expert Content Validation

Qualitative study of open-ended responses by experts identified a number of themes in need of attention. Three experts wrote comments on the open-ended questions in Instrument 2. Of criteria to add / delete, one expert said, '*The Collaboration requirement is problematic in that not all PBL assignments in our department are group projects. Many students do individual research projects / teaching demonstrations. The rubric should specify if this criterion is mandatory or optional by project type.'* Further expert comments: *'I would add a criterion for 'Use of sources and referencing. In my experience, undergraduate students in ELT departments have difficulties with correct referencing and avoiding plagiarism in their research papers. This is a part of PBL that is lacking in the present rubric.'* Regarding the language of performance descriptors, one expert said: *'There is some confusion as to the difference between 'Proficient' and 'Exemplary' for Critical Thinking. 'Provides reasoned analysis with evidence' vs 'Evaluates diverse perspectives,' it is overlap. I would differentiate the Exemplary level more by requiring that the candidate 'synthesizes information from multiple sources.'* All three experts agreed that the differentiation alternatives (Option A, B and C) were realistic and appropriate for the Indonesian environment. However, one expert said: *'Option C (varying criteria weights for different student preparedness levels) is theoretical but lecturers will require examples of how to do this. Many lecturers will default to Option A if not trained. My suggestion is to include a teacher handbook with sample scenarios.'*

Revisions Made Based on Expert Feedback

Three changes to the DAR-PBL were made based on quantitative and qualitative content validity testing results before proceeding into subsequent phases. Three changes were made: The Collaboration criterion was updated to separate descriptors for individual and group project contexts. This criteria was re-averaged around 'Independent Work Ethic' in which descriptors focus on completing tasks, managing time effectively and learned from self-directed learning for individual projects separately. The original collaboration descriptors were preserved as they applied to group projects. Second, a seventh criterion, called 'Use of Sources and Referencing', was added for research-based PBL projects with performance levels from 'No citation of sources' (Emerging) to 'Consistent, accurate APA-style citation of references that are relevant to the arguments presented' (Exemplary). Third, a one-page teacher guide with examples of implementing Option C (different weights for criteria based on readiness) was created, including sample scenarios summarizing low-, medium- and high-readiness students.

Face Validity

The Face Validity was measured by the Face Validity, Construct Validity and Usability Questionnaire (Instrument 3, Part A and Part C) that was administered to six lecturers and 20 undergraduate students after they reviewed and used the DAR-PBL. The lecturers had a minimum of 2 years of teaching experience in the department of English Language Education and had recently implemented PBL in their courses. The students were in 3rd and 5th semester of study and had taken at least one English course based on PBL. After a brief orientation of 15 minutes, all participants were asked to review the DAR-PBL individually. The lecturers then scored three example student work samples using the rubric to imitate actual use before filling in the questionnaire. All items were measured on a 5-point scale (1 = strongly disagree to 5 = strongly agree). Acceptable face validity was indicated by a mean score of ≥ 4.0 on the 5-point scale.

Table 3 shows the mean scores and standard deviations of all the face validity items (Parts A and C) for lecturers and students separately.

Table 3. Face Validity Mean Scores (Clarity and Fairness)

Item	Lecturers (n=6) Mean (SD)	Students (n=20) Mean (SD)
Part A: Clarity		
A1: Well defined rubrics criteria and easily understandable	4.67 (0.52)	4.45 (0.69)
A2: The performance level descriptors (Emerging, Developing, Proficient, Exemplary) are distinct from each other	4.50 (0.55)	4.30 (0.73)
A3: The wording of the rubric is adequate for university level English courses	4.83 (0.41)	4.60 (0.60)
Part C: Fairness		
C1: The rubric provides all students with a fair opportunity to show what they know	4.33 (0.82)	4.25 (0.79)
C2: The options for differentiation (A, B or C) are possible and doable	4.50 (0.55)	4.40 (0.68)
C3: The rubric does not favour any certain learning style over others	4.17 (0.75)	4.10 (0.85)
Face Validity Mean	**4.50 (0.60)**	**4.35 (0.72)**

As can be seen in Table 3, for both lecturers and students, all mean scores for clarity and fairness were above the required level of 4.0. Lecturers assessed the language of the rubric as most appropriate (A3: M = 4.83, SD = 0.41), with five of the six lecturers strongly agreeing and one concurring. This item was also rated highly by students (M = 4.60, SD = 0.60), with twelve students strongly agreeing, six agreeing and two neutral. Both groups evaluated fairness about learning styles the lowest (C3: lecturers M = 4.17, students M = 4.10). For this issue two instructors agreed and four strongly agreed, for students eight strongly agreed, nine agreed, two were neutral and one disagreed. The face validity mean was 4.50 for lecturers and 4.35 for students.

Qualitative Findings from Face Validity Questionnaire

The open-ended responses from part F of the questionnaire were qualitatively analzed and some themes emerged across both the lecturers and students. The four open-ended questions were answered in writing by 26 participants (6 lecturers and 20 students).

Theme 1: Appreciation for Clarity and Specificity

The noted good part of the rubric was the clarity and specificity of the performance level descriptors. One speaker commented, *'The difference between 'Developing' and 'Proficient' is very evident, as opposed to many rubrics that I have used before. I can see exactly what a student needs to accomplish to progress from one level to the next.'* One student wrote: *'This rubric is quite simple. I know exactly what I need to accomplish to earn a score of Proficient / Exemplary. Other rubrics I have tried were confusing with adjectives like 'good' / 'satisfactory'.'*

Theme 2: Reduced Assessment Anxiety

Several students said that the rubric lessened their concern about testing because they could understand exactly what to expect at each level. As one student put it, *'I always get scared when my lecturer marks my project since I don't know what they are looking for. This rubric helps me to assess my work against the standards before I submit. That makes me feel less anxious.'* Another student said: *'The distinction options are useful since I can select which criteria to focus on. I am strong at subject knowledge. Option B allows me display my strength.'*

Theme 3: Initial Difficulty with Scoring Formula

Three of the six lecturers reported that the scoring formula was initially confusing, but after practicing with the sample work, it was easy to use. One lecturer put it this way: *'At first I was perplexed about how to calculate the weighted raw score. The formula appeared complex. But once I graded the three sample projects, I got it. I think lecturers need practice time before applying this rubric for grades.'* A second lecturer suggested: *'Maybe put in a calculator / spreadsheet template so lecturers don't have to conduct the calculation themselves. This would save time and reduce errors.'*

Theme 4: Request for Concrete Examples

Two students and one lecturer recommended including more examples of what each performance level means in

practice, particularly for the Creativity and Originality criterion. One student wrote: '*I don't really understand what 'innovative' / 'transformative' means in practice, for Creativity. Can you give examples of student work that earned each level?'* A lecturer said: *'It would be quite beneficial to have an annotated exemplar for every criterion and level to instruct new lecturers who will be using this rubric.'*

Theme 5: Positive Reaction to Differentiation Options

All six lecturers and fifteen out of twenty students remarked positively on the differentiation possibilities. One lecturer said, *'Option C is just right for my varied ability classes. I have students who are really good in speaking but poor in writing. 'It's fair since we can modify weights for various students and it fits with concepts of individualized instruction.'* One student commented, *'I like Option B more because I am not excellent at working in groups. I like to work alone. The ability to select 5 out of 6 criteria means I can skip Collaboration and still get a decent score.'*

Construct Validity

Part D of the Face Validity, Construct Validity, and Usability Questionnaire (Instrument 3) examined perceived usefulness of the DAR-PBL and provided evidence of construct validity. The same 6 lecturers and 20 students participated in this part. All items were assessed on a 5-point scale (1= strongly disagree to 5= strongly agree). A mean score of ≥ 4.0 was considered as acceptable evidence of construct validity as greater perceived utility supports the idea that the rubric measures what it promises to assess. Table 4 shows the means and standard deviations of all items of construct validity (Part D) for the lecturers and students separately.

Table 4. Construct Validity Mean Scores (Perceived Usefulness)

Item	Lecturers (n=6) Mean (SD)	Students (n=20) Mean (SD)
D1: This rubric effectively analyses the students' learning from the PBL project.	4.33 (0.82)	4.20 (0.77)
D2: Rubric provides feedback to support student learning	4.67 (0.52)	4.50 (0.69)
D3: I would encourage this rubric to be implemented in other English Language Education departments	4.83 (0.41)	4.55 (0.60)
Construct Validity Mean	**4.61 (0.58)**	**4.42 (0.69)**

As seen in Table 4, the three items had a score over the acceptable level of 4.0, both for lecturers and students. The highest score from both groups was to promote the rubric to other departments (D3: lecturers M = 4.83, students M = 4.55). Five lecturers highly agreed and one agreed with this remark. Among the students, 13 strongly agreed, 6 agreed, and 1 was impartial. This shows a great degree of acceptance of the instrument and show that they find it of value beyond the local context. The second highest rated criterion was the rubric's capacity to deliver relevant feedback (D2) (lecturers M = 4.67, students M = 4.50). In the open-ended portion one speaker noted: *'The descriptors are so specific that I can just circle the level and return the rubric to the student. They can see exactly what they need to work on without me writing lengthy comments.'* The lowest result among the three items of construct validity was the item evaluating accuracy of measurement (D1), but still over the threshold (lecturers M = 4.33, students M = 4.20). D1 was agreed by 2 lecturers and strongly agreed by 4 lecturers. Students: 10 highly agreed, 8 agreed, 2 neutral. One student who assessed D1 as neutral said, *'I believe the rubric measures what I have learned, but I am not 100% sure because this is my first time using it. Maybe after utilizing it on more projects I will have a better idea.'*

Comparison Between 3rd and 5th Semester Students

Perceived usefulness for students with different academic experiences was also investigated by splitting up 20 students into two groups, 3rd semester students (n=10) and 5th semester students (n=10). An independent samples t-test was used to compare the mean scores of construct

validity between two groups. Table 5 shows the comparison of concept validity evaluations between 3rd and 5th semester students.

Table 5. Construct Validity Mean Scores by Student Semester

Student Group	n	Mean (SD)	t-value	p-value
3rd Semester	10	4.40 (0.71)	-0.24	0.81
5th Semester	10	4.43 (0.68)		

As can be seen in Table 5, there was no statistically significant difference between 3rd semester students (M = 4.40, SD = 0.71), and 5th semester students (M = 4.43, SD = 0.68), t(18) = -0.24, p = 0.81. This shows that perceived usefulness was consistent across different levels of academic expertise.

Relationship Between Face Validity and Construct Validity

A Pearson correlation coefficient was calculated to determine the relationship between the total face validity scores (Parts A and C) and the construct validity scores (Part D) for all the 26 participants (6 lecturers + 20 students). There was a significant positive association, r(24) = 0.78, p < 0.001, showing that participants who rated the rubric higher on clarity and fairness also rated it higher on perceived usefulness. This gives further evidence of the construct validity of the DAR-PBL in that the transparency of the rubric is linked to the users' perception that it measures what it purports to measure.

Inter-Rater Reliability

The inter-rater reliability was determined by six lecturers independently scoring 20 authentic PBL student work examples with the final revised version of the DAR-PBL (Instrument 4). The 20 work samples were selected from the courses of the participating lecturers. They included several sorts of projects: portfolios (n=5), research posters (n=5), teaching videos (n=5), and lesson plans (n=5). All 20 work samples were graded by the six lecturers during a two-week period. To reduce order effects, the 20 work samples were presented in random order to each rater. The Intraclass Correlation Coefficient (ICC) was calculated based on a two-way random effects model for absolute agreement (ICC 2,1) as suggested by Koo & Li (2016). Moreover, Fleiss' Kappa was calculated for category agreement on performance levels (Emerging, Developing, Proficient, Exemplary) to offer a more strict measure of rater agreement beyond absolute scores. Table 6 shows the inter-rater reliability for the final ratings for the 20 work samples for all six raters.

Table 6. Inter-Rater Reliability Results

Measure	Value	95% Confidence Interval	Interpretation
ICC (2,1) for final scores	0.87	0.79 – 0.93	Good reliability
Fleiss' Kappa (categorical)	0.68	0.61 – 0.75	Substantial agreement

The ICC for the final scores, based on all six raters, was 0.87 (95% CI: 0.79 to 0.93), showing good reliability (Koo & Li, 2016) (see Table 6). There was significant agreement beyond chance on categorical agreement on performance levels (Fleiss' Kappa = 0.68, p < 0.001). The 95% confidence interval for Fleiss' Kappa was between 0.61 and 0.75, indicating that the true value in the population lies in the area of high agreement.

Inter-Rater Reliability by Individual Criterion

To determine which rubric criterion contributed the most and least to overall rater agreement, ICC values were calculated for each of the six criteria across all six raters and 20 work samples. The ICC values for each criterion are shown in Table 7.

Table 7. ICC Values for Individual Criteria

Criterion	ICC (2,1)	95% CI	Interpretation
Content Knowledge	0.91	0.85 – 0.95	Excellent
English Language Use	0.89	0.82 – 0.94	Good
Critical Thinking	0.84	0.76 – 0.90	Good
Collaboration	0.76	0.66 – 0.84	Good
Creativity & Originality	0.71	0.60 – 0.80	Moderate
Organization & Presentation	0.85	0.77 – 0.91	Good

The highest rater agreement was found for Content Knowledge, ICC = 0.91, excellent, as shown in Table 7. This means that the lecturers were very consistent in their assessment of the correctness and breadth of the content knowledge shown in the students' PBL projects. Good reliability was also found for English Language Use (ICC = 0.89) and Organization & Presentation (ICC = 0.85).

Critical Thinking (ICC = 0.84) and Collaboration (ICC = 0.76) showed good reliability, though Collaboration was at the lower end of the good range. The lowest agreement was found for Creativity & Originality (ICC = 0.71, moderate). This indicates that the lecturers showed greater variation in their assessment of student creativity and originality than other factors.

Inter-Rater Reliability by Project Type

It is also computed ICC values separately for each of the four project categories (portfolios, research posters, instructional videos, and lesson plans) to determine whether rater agreement was dependent on the kind of PBL work sample. Table 8 displays the ICC values for final scores by project type.

Table 8. ICC Values by Project Type

Project Type	Number of Samples	ICC (2,1)	95% CI	Interpretation
Portfolios	5	0.91	0.83 – 0.96	Excellent
Research Posters	5	0.85	0.76 – 0.92	Good
Teaching Videos	5	0.82	0.72 – 0.89	Good
Lesson Plans	5	0.88	0.80 – 0.94	Good

The highest rater agreement, as shown in Table 8, was for portfolios (ICC = 0.91, outstanding), followed by lesson plans (ICC = 0.88) and research posters (ICC = 0.85). Teaching videos had the lowest agreement (ICC = 0.82) among the four categories of projects, but this is still in the region of high reliability. The lower agreement for teaching videos can be due to the greater subjectivity in evaluating oral communication and classroom management skills when compared to written goods like portfolios and lesson plans.

Distribution of Performance Level Assignments

To evaluate how the six raters dispersed their scoring across the four performance levels, the frequency of the six raters' scoring across the four performance levels was determined over all 20 work samples and six raters (120 total ratings). Table 9 shows the frequency and percentage of each level of performance assigned by all raters.

Table 9. Distribution of Performance Level Assignments (All Raters, All Samples)

Performance Level	Frequency	Percentage
Emerging (1)	18	15.0%
Developing (2)	42	35.0%
Proficient (3)	45	37.5%
Exemplary (4)	15	12.5%
Total	**120**	**100%**

As seen in Table 9, most of the ratings were in the Developing (35.0%) and Proficient (37.5%) levels. This is expected as the work samples came from undergraduate students in an English Language Education department. Emerging (15.0%) and Exemplary (12.5%) were less commonly allocated, showing that the rubric discriminate between lower and higher performing students without too much grouping at the extremes.

Agreement on Exact Performance Level

Furthermore, to provide a more complete information of the rater agreement, the percentage of pairs of work samples where all six raters agreed exactly on the level of performance was computed per criterion. Table 10 shows the percentage of work samples (out of 20) for which all six raters gave the same exact performance rating for each criterion.

Table 10. Perfect Agreement (6 of 6 Raters) by Criterion

Criterion	Number of Samples with Perfect Agreement	Percentage
Content Knowledge	12	60%
English Language Use	10	50%
Critical Thinking	8	40%
Collaboration	7	35%
Creativity & Originality	5	25%
Organization & Presentation	9	45%

As shown in Table 10, Content Knowledge had the highest percentage of perfect agreement of all six raters (60% of work samples), showing this criterion was the most objective and consistently implemented. The criterion Creativity & Originality showed the lowest perfect agreement (25% of work samples) further confirming the notion that this criterion was the most difficult for raters to apply consistently. For the other 75% of work samples on the Creativity criterion, at least one rater was off by one performance level from the others.

Rater Consistency Across Time

To test the consistency of individual raters across the two-week period, six raters rescored a subset of five work samples

(randomly selected) one week following the original rating. Test-retest reliability was calculated using Pearson correlation between the first and second scores for each rater. The test-retest correlation coefficient for each individual rater is shown in Table 11.

Table 11. Test-Retest Reliability by Individual Rater

Rater ID	Correlation (r)	Interpretation
R01	0.94	Excellent consistency
R02	0.91	Excellent consistency
R03	0.89	Good consistency
R04	0.92	Excellent consistency
R05	0.88	Good consistency
R06	0.90	Excellent consistency
Mean	**0.91**	**Excellent consistency**

The results are written in Table 11; all six raters scored their first and second sessions very consistently with individual correlations between scores from 0.88 to 0.94 (mean r = 0.91). This suggests that the rubric yielded consistent scores throughout time and that the two-week scoring period did not contribute to rater tiredness or divergence in scoring standards.

Discussion

The results of this research offered empirical evidence for the validity and reliability of the Differentiated Assessment Rubric for Project-Based Learning (DAR-PBL) among all Indonesian departments that are predominantly focused on English Language Education. Five out of the six rubric criteria, namely Presentation Format, English Language Use, Content Representation & Originality, Connection and Convergence showed fair to good agreement (Aiken's V = 0.75–0.90), while English Language Use reached perfect agreement (V=1.00). Such content validity supports the tenet of assessment design in differentiated instruction that criteria must truly reflect the construct being measured while reflecting sensitivity to learner diversity (Tomlinson et al., 2013). Although the lower coefficient (V = 0.67) for the Collaboration criterion was initially alarming, qualitative feedback from experts found that inapplicability of this criterion was not due to irrelevance, but rather its irrelevance to individual projects. This contrasts with a project typically being one structure across various student rubrics, an assumption present in the work of (Brookhart & Nitko, 2018) where they pose that rubric criteria must have flexibility around various task shapes to align with different assessment contexts. Restructuring the Collaboration criterion into descriptors for individual and group assessments shows an interpretation of authenticity in performance-based assessment (Ajjawi et al., 2024).

The mean scores for lecturers (M = 4.50) and students (M = 4.35), were all above 4, with the highest ratings suggesting that users found the DAR-PBL to be clear, fair, and relevant to university-level English courses. These results are important as face validity, trivialized into the realm of seeming goodness-of-fit, but recently reaffirmed its status, core criterion of assessment usability, including in linguistically / culturally diverse contexts. The qualitative theme of students reporting decreased assessment anxiety speaks directly to the theoretical claim that transparent, differentiated rubrics lower affective barriers to performance and thus facilitate learners being at their most complete measure of true competencies (Panadero & Jonsson, 2020). The response in comparison of differentiation options among lecturers and students is consistent with the wider literature on Universal Design for Learning (UDL), which argues that while providing multiple means to show competence, demonstrates increased both equity and validity (Capp, 2020). This is further supported by the strong positive correlation between face validity and construct validity ($r=0.78$, $p<0.001$) suggesting that when users perceive a rubric as clear and fair they also agree it measures what it claims to measure, findings consistent with Kane's (2016) framework of validity in classroom-based assessment contexts.

The construct validity evidence, based on the ratings of perceived usefulness (M = 4.61 lectures; M = 4.42 students), shows

that this instrument can be working properly as designed with a strong support for the instrument to be non-devices of shortcomings for punctual preparation assessing in DAR-PBL classroom sprightly running. Recommendation to other departments, which was rated the highest of any item indicating that users viewed the instrument as having broader applicability than just locally. This result has based on a methodological lesson, that is putting aside the multitrait-multimethod approach to construct validation and subsequently embracing a user-centered perspective hailed by Zumbo & Chan, (2014) which states that perceived consequences and usefulness are key for validity arguments in educational measurement. The lack of a statistically significant difference ($p = 0.81$) between semesters indicates stable perceived utility across levels of academic experience, which is needed for longitudinal implementation of the rubric (Schuwirth & van der Vleuten, 2020). In addition, the qualitative finding that supports the utility of rubrics for delivering granular feedback without detailed written feedback in turn addresses a recognised barrier to PBL scale-up, the resource demands and assessment burden have been shown to undermine the intention to teach with PBL at scale.

Results reveal that inter-rater reliabilities (ICC = 0.87, Fleiss' Kappa = 0.68) showed good to substantial agreement between six independent raters assessing twenty authentic PBL artifacts. These values adhere to / exceed the recommended benchmarks for high-stakes classroom assessment in applied linguistics contexts (Plonsky & Derrick, 2016). The fact that Content Knowledge produced the highest ICC (0.91) and Creativity & Originality produced the lowest (0.71, moderate) at the criterion level complements the accepted principle governing rater-mediated assessment that constructs dependent on subjective context yield greater variability among raters (Wind & Peterson, 2018). Still, one would not consider the moderate reliability for Creativity as a strike against. Rather, it illustrates a well-known measurement issue in PBL assessment, namely, that creativity is often operationalized differently by evaluators on the basis of their disciplinary and cultural backgrounds (Cropley & Patston, 2019). The relatively high reliability for the Creativity criterion (ICC = 0.71) is rather encouraging, given the inherent subjectivity of this construct and reports of considerably lower inter-rater agreement for creativity-related constructs with non-differentiated rubrics in Asian EFL contexts (Kim, Hyunah, 2016). The differences in ICC by project mode where the agreement for teaching videos was lower (0.82) than for portfolios (0.91), are consistent with performance assessment literature indicating multimodal artifacts require more rater training compared to text-based products (Gomes & Jelihovschi, 2020).

In addition to the positive rating performance, it is also noted excellent inter-rater consistency for test-retest reliability (mean $r = 0.91$) over 2-week period, which suggests that the rubric descriptors were stable enough to be resistant to rater drift–a well-documented threat to reliable scoring for performance-based assessment (Myford & Wolfe, 2003). This stability is all the more crucial for differentiated features of the rubric (Options A, B and C) as a lack of rater consistency across time would sabotage one of the main functions. This allows flexible weighting and criterion selection that are offered by these innovative approaches to scoring (Barua & Lockee, 2025). The 15% of performance level assignments in the Emerging category and only 12.5% in Exemplary indicates that the rubric effectively discriminates between student performance without ceiling or floor effects. The logic of optimal test discrimination refers to the idea that most respondents will score in the mid-level scale categories where variance is most

meaningful (Shermis, 2007). Discovery that 60% of work samples achieved six-rater perfect agreement on Content Knowledge versus just 25% on Creativity & Originality provides a direction for future rater training, differentiated rubrics require calibration exercises to support more subjective rubric criteria (Gunnell et al., 2016).

These validity and reliability results provide three contributions to the literature on differentiated assessment. This study establishes the empirical basis for introducing differentiation principles, namely, flexible criterion weighting and student choice of assessment focus, into a rubric while maintaining its psychometric quality (Goyibova et al., 2025). Secondly, the successful validation of the DAR-PBL in an Indonesian EFL context contradicts a widely held view that standardized, one-size-fits-all rubrics are type by nature more reliable than differentiated alternatives (DeLuca et al., 2012). Third, this study utilized a replication of validation procedures for developing objective grading instruments by blending quantitative measures (Aiken's V, ICC, Fleiss' Kappa) with subsequent qualitative thematic analysis of expert judges, lecturers and students responses (Chapelle, 2022). The validation of the DAR-PBL, as a test intended for use in English Language Education departments in Indonesian higher education provides an anchor from which cross-contextual adaptations can be compared (Knoch & Macqueen, 2019). Future study should assess the predictive validity of this rubric by analyzing DAR-PBL scores and related academic success and examine if differentiated rubrics directly create more equitable behavior for learners of various language and education backgrounds.

Conclusion

The present study aimed to create and validate the Differentiated Assessment Rubric for Project-Based Learning (DAR-PBL) because of the lack of empirically-based assessment tools applicable for diversity in learners' contexts in Indonesian English Language Education departments. The research question examined the level of which newly developed DAR-PBL shows acceptable content validity, face validity, construct validity and inter-rater reliability to assess project-based learning competencies in this context. The results provide strong support that the DAR-PBL satisfies or surpasses accepted psychometric criteria across each of the four dimensions of validity and reliability assessed. In terms of content validity, results suggested that Aiken's V coefficients for five of the six rubric criteria were above the acceptable range (≥ 0.75), with perfect agreement ($V=1.00$) across the three expert judges on English Language Use. Specifically, the Collaboration criterion started off just below the lower threshold ($V = 0.67$) but through qualitative expert feedback was revised significantly to become a far more flexible criterion with different descriptors for both individual and group project contexts, along with an entirely new seventh criterion that addressed source use and referencing. And the unanimous agreement of the expert panel on rubric functional suitability for Indonesian ELT settings (mean = 4.00) provide additional evidence for content validity. The results of this research confirm that the DAR-PBL has indicators that are adequate, representative and comprehensive to measure learning outcomes about the application of PBL in higher education in Indonesia as well as shown by expert input which is needed when identifying or correcting indicators which do not have equivalent between projects.

The evidence for face validity was strong, with means from both lecturers ($M = 4.50$) and students ($M = 4.35$) exceeding the established minimum level of acceptability of 4.0 on a scale from 1-5. Rubric language clarity was rated highly by lecturers ($M = 4.83$), and students were in strong agreement that the rubric was fair

and transparent. Qualitative thematic analysis resulted in five prominent themes. They are appreciation of clarity and specificity of performance descriptors, reduced assessment anxiety for students, initial difficulty with the scoring formula that became easier with practice, requests for explicit examples of performance levels, and universally positive allusions to differentiation options. These results suggest that although the surface validity of the DAR-PBL was established through relevant documentation, participants perceive it as having clarity and fairness, and being appropriate for purpose. That students self-reported less anxiety in performing with the rubric advocates a significant affective function of differentiated assessment tools beyond their measurement purpose. The construct validity, measured through perceived usefulness ratings, was well supported with mean scores of 4.61 and 4.42 for lecturers and students respectively. The top-rated item addressed the willingness to recommend rubric use in other English Language Education departments (D3: lecturers M = 4.83, students M = 4.55), signalling that users believed that the instrument offers value outside of this specific research context. A Cumulative Score on the Rubric for providing examples of Descriptors and giving meaningful feedback (anchors D2: lecturers M = 4.67, students M = 4.50) was also high with one lecturer commenting that the specific descriptors meant that lengthy written comments were unnecessary Graduates Classroom Study Results The statistically non-significant difference between third-semester and fifth-semester students ($p = 0.81$) illustrates that perceived usefulness is invariant across levels of academic experience. In addition, there was a strong positive correlation between face and construct validity scores ($r = 0.78$, $p < 0.001$), which converges with the finding that mentors who rated rubric items as clear and fair also indicated that the rubric measured what it was intended to measure (thus supporting the validity argument).

The inter-rater reliability analysis produced an Intraclass Correlation Coefficient for final scores of 0.87 (indicating good reliability based on established benchmarks, also with a Fleiss' Kappa coefficient of 0.68 representing substantial agreement beyond chance). In an analysis by individual criterion, Content Knowledge had the highest reliability (ICC = 0.91, excellent), followed by Creativity and Originality with moderate reliability (ICC = 0.71), which speaks to how much more subjective creative work is compared to other areas of study. Critically, test-retest reliability across a two-week period was excellent (mean r = 0.91), showing that raters adhered to the same standards over time and suggesting the rubric is resistant to rater drift. These reliability findings support that the DAR-PBL can be employed across multiple lecturers consistently and yield reproducible scores over time—two key attributes for any assessment tool used to make high-stakes decisions regarding student learning. This study has a number of limitations that should be noted. This study has limitations in that the sample of expert judges (n=3), lecturers (n=6) and student work samples (n=20) was relatively small, which also limits possibly the generalisability of findings to field although these are consistent with instrument validation studies in rubric development literature. Second, the study was performed at a small number of institutions in Indonesia, while the rubric is targeted to Indonesian English Language Education departments and shows good psychometric properties for this context, generalization beyond this setting would benefit from further validation across different parts of the country as well as institutional type. Third, the current validation only looked at six core criteria plus the additional seventh criterion regarding source use, but future research should explore other criteria that

might apply to particular types of PBL projects. Fourth, evidence of construct validity was based mainly on perceived usefulness ratings and future research should link scores from DAR-PBL with other well-validated measures of PBL.

In theory, the validated DAR-PBL validation successfully complements both differentiated instruction theory and can extend into PBL evaluation providing empirical evidence that differentiation principles can be realized in a rubric form without compromising psychometric quality. In practice, the DAR-PBL provides Indonesian ELT lecturers with an objective and research-based PBL assessment tool to be fairer, as well as more transparent and credible in judging students' performance by providing guidance for students' learning articulation tailored to their profiles. Rubrics are designed for a standard but classroom diversity cannot be contained in one size fits all rubrics, therefore differentiation options embedded into the rubric itself including adjustable criterion weightings, student selection of assessment focus and contextual adaptations, such as individual versus group work (less-weighted in favour of collaboration). A teacher guide accompanying the textbook with different scenarios of implementing option C (varying criteria weights based on readiness) addresses expert comments about actionability by acknowledging that even strong rubrics require training and resource support to help teachers implement this innovation effectively.

References

Abdullah, M. N. L. Y. (2020). Student-centered philosophies and policy developments in Asian higher education. In *The Routledge International Handbook of Student-Centered Learning and Teaching in Higher Education*.

Adnan, Suwandi, S., Nurkamto, J., & Setiawan, B. (2019). Teacher competence in authentic and integrative assessment in Indonesian language learning. *International Journal of Instruction*, *12*(1). https://doi.org/10.29333/iji.2019.12145a

Aiken, L. R. (1985). Three coefficients for analyzing the reliability and validity of ratings. *Educational and Psychological Measurement*, *45*(1). https://doi.org/10.1177/0013164485451012

Ajjawi, R., Tai, J., Dollinger, M., Dawson, P., Boud, D., & Bearman, M. (2024). From authentic assessment to authenticity in assessment: broadening perspectives. *Assessment and Evaluation in Higher Education*, *49*(4). https://doi.org/10.1080/02602938.2023.2271193

Barua, L., & Lockee, B. (2025). Flexible Assessment in Higher Education: A Comprehensive Review of Strategies and Implications. *TechTrends*, *69*(2). https://doi.org/10.1007/s11528-025-01039-3

Brookhart, S. M. (2018). Appropriate Criteria: Key to Effective Rubrics. In *Frontiers in Education* (Vol. 3). https://doi.org/10.3389/feduc.2018.00022

Brookhart, S. M., & Nitko, A. J. (2018). Educational assessment of students. *Human Movement Science*, *53*(1).

Capp, M. J. (2020). Teacher confidence to implement the principles, guidelines, and checkpoints of universal design for learning*. *International Journal of Inclusive Education*, *24*(7). https://doi.org/10.1080/13603116.2018.1482014

Chapelle, C. A. (2022). Argument-Based Validation in Testing and Assessment. In *Argument-Based Validation in Testing and Assessment*.

https://doi.org/10.4135/9781071878811

Cropley, D. H., & Patston, T. J. (2019). *Supporting Creative Teaching and Learning in the Classroom: Myths, Models, and Measures*. https://doi.org/10.1007/978-3-319-90272-2_15

DeLuca, C., Luu, K., Sun, Y., & Klinger, D. A. (2012). Assessment for learning in the classroom: Barriers to implementation and possibilities for teacher professional learning. *Assessment Matters*, *4*. https://doi.org/10.18296/am.0104

DeVellis, R. F. (2017). Scale Development Theory and Applications (Fourth Edition). *SAGE Publication*, *4*.

Gomes, C. M. A., & Jelihovschi, E. (2020). Presenting the Regression Tree Method and its application in a large-scale educational dataset. *International Journal of Research and Method in Education*, *43*(2). https://doi.org/10.1080/1743727X.2019.1654992

Goodwin, L. D., & Leech, N. L. (2003). The Meaning of Validity in the New Standards for Educational and Psychological Testing: Implications for Measurement Courses. In *Measurement and Evaluation in Counseling and Development* (Vol. 36, Number 3). https://doi.org/10.1080/07481756.2003.11909741

Goyibova, N., Muslimov, N., Sabirova, G., Kadirova, N., & Samatova, B. (2025). Differentiation approach in education: Tailoring instruction for diverse learner needs. In *MethodsX* (Vol. 14). https://doi.org/10.1016/j.mex.2025.103163

Gunnell, K. L., Fowler, D., & Colaizzi, K. (2016). Inter-rater reliability calibration program: critical components for competency-based education. *The Journal of Competency-Based Education*, *1*(1). https://doi.org/10.1002/cbe2.1010

Hall, T., Vue, G., Strangman, N., & Meyer, A. (2004). Differentiated instruction and implications for UDL implementation (effective classroom practices report). *National Center on Accessing the General Curriculum*, *4*(2).

Hasanah, E., Suyatno, S., Maryani, I., Badar, M. I. Al, Fitria, Y., & Patmasari, L. (2022). Conceptual Model of Differentiated-Instruction (DI) Based on Teachers' Experiences in Indonesia. *Education Sciences*, *12*(10). https://doi.org/10.3390/educsci12100650

Hustarna, Fitri, E. N., Efriza, D., & Arif, N. (2024). EFL students' perception toward the use of Instagram for academic speaking skill learning. *Research and Innovation in Language Learning*, *7*(1).

Kane, M. T. (2016). Explicating validity. *Assessment in Education: Principles, Policy and Practice*, *23*(2). https://doi.org/10.1080/0969594X.2015.1060192

Kim, Hyunah. (2016). Comparing Native and Non-native Rater Assessments of Korean Oral Proficiency: A FACETS analysis. *Korean Language Education Research*, *51*(5). https://doi.org/10.20880/kler.2016.51.5.83

Knoch, U., & Macqueen, S. (2019). Assessing english for professional purposes. In *Assessing English for Professional Purposes*. https://doi.org/10.4324/9780429340383

Koo, T. K., & Li, M. Y. (2016). A Guideline of Selecting and Reporting Intraclass Correlation Coefficients for Reliability Research. *Journal of Chiropractic Medicine*, *15*(2).

https://doi.org/10.1016/j.jcm.2016.02.012

Kupchyk, L., & Litvinchuk, A. (2020). Differentiated instruction in English learning, teaching and assessment in non-language universities12. *Advanced Education*, *2020*(15). https://doi.org/10.20535/2410-8286.168585

Larmer, J., Mergendoller, J. R., & Boss, S. (2015). Setting the standard for project based learning. In *Engineering* (Numbers 1–2).

Myford, C. M., & Wolfe, E. W. (2003). Detecting and Measuring Rater Effects Using Many-Facet Rasch Measurement: Part I. *Journal of Applied Measurement*, *4*(4).

Panadero, E., & Jonsson, A. (2020). A critical review of the arguments against the use of rubrics. In *Educational Research Review* (Vol. 30). https://doi.org/10.1016/j.edurev.2020.100329

Phuong, H. Y., Phan, Q. T., & Le, T. T. (2023). The effects of using analytical rubrics in peer and self-assessment on EFL students' writing proficiency: a Vietnamese contextual study. *Language Testing in Asia*, *13*(1). https://doi.org/10.1186/s40468-023-00256-y

Plonsky, L., & Derrick, D. J. (2016). A Meta-Analysis of Reliability Coefficients in Second Language Research. *Modern Language Journal*, *100*(2). https://doi.org/10.1111/modl.12335

Religioni, A., Meilinda, C., Alawiyah, A., & Farrel, L. M. (2024). Differentiated Instructions in ESL and EFL Classrooms: A Systematic Literature Review. *STAIRS: English Language Education Journal*, *5*(2). https://doi.org/10.21009/stairs.5.2.5

Schuwirth, L. W. T., & van der Vleuten, C. P. M. (2020). A history of assessment in medical education. *Advances in Health Sciences Education*, *25*(5). https://doi.org/10.1007/s10459-020-10003-0

Shermis, M. D. (2007). A Review of "Modern Measurement: Theory, Principles, and Applications of Mental Appraisal." *International Journal of Testing*, *7*(4). https://doi.org/10.1080/15305050701632288

Sun, Y., & Xiao, L. (2024). Research trends and hotspots of differentiated instruction over the past two decades (2000-2020): a bibliometric analysis. In *Educational Studies* (Vol. 50, Number 2). https://doi.org/10.1080/03055698.2021.1937945

Thornhill-Miller, B., Camarda, A., Mercier, M., Burkhardt, J. M., Morisseau, T., Bourgeois-Bougrine, S., Vinchon, F., El Hayek, S., Augereau-Landais, M., Mourey, F., Feybesse, C., Sundquist, D., & Lubart, T. (2023). Creativity, Critical Thinking, Communication, and Collaboration: Assessment, Certification, and Promotion of 21st Century Skills for the Future of Work and Education. In *Journal of Intelligence* (Vol. 11, Number 3). https://doi.org/10.3390/jintelligence11030054

Tomlinson, C. A. (2001). How TO Differentiate instruction in mixed-ability classrooms. In *Association for Supervision and Curriculum Development*.

Tomlinson, C. A. (2017). How to Differentiate Instruction in Academically Diverse Classrooms. *Differentiate Instruction: In Academically Diverse Classooms*.

Tomlinson, C. A., Moon, T., & Imbeau, M. B. (2013). Assessment and Student Success in A Differentiated Classroom. In *Association for Supervision and Curriculum Development*.

Wind, S. A., & Peterson, M. E. (2018). A systematic review of methods for evaluating rating quality in language assessment. *Language Testing*, *35*(2). https://doi.org/10.1177/0265532216686999

Zumbo, B. D., & Chan, E. K. H. (2014). Setting the stage for validity and validation in social, behavioral, and health sciences: Trends in validation practices. In *Social Indicators Research Series* (Vol. 54). https://doi.org/10.1007/978-3-319-07794-9_1