ORIGINAL RESEARCH

Multistakeholder Assessment of Project-Based Service-Learning in Medical Education: A Comparative Evaluation

Shih-Chieh Liao[1], Yueh-Nu Hung[2], Chia-Rung Chang [ID][1], You-Xin Ting [ID][1]

[1]School of Medicine, China Medical University, Taichung City, Taiwan; [2]Department of English, National Taichung University of Education, Taichung City, Taiwan

Correspondence: Yueh-Nu Hung, Department of English, National Taichung University of Education, 140 min-Sheng Road, Taichung, 403, Taiwan, Tel +886 4 22183845, Fax +886 4 22183460, Email yuehnuhung@gm.ntcu.edu.tw

Purpose: Traditional single-assessment models in service-learning courses do not facilitate comprehensive assessments of learning outcomes. Effective assessments should incorporate perspectives from multiple stakeholders. The present study developed a service-learning course assessment model that incorporates assessments from multiple stakeholders, compared assessments between stakeholder types, and explored the effects of evaluator–student relationship.

Participants and Methods: The study recruited 126 students from a service-learning course at China Medical University in 2024. Six different groups of stakeholders, namely peers, teaching assistants, service institutions, primary instructors, group instructors, and final report evaluators, evaluated student performance and learning outcomes. Experts ensured that assessment criteria were relevant and comprehensive. Confirmatory factor and principal component analyses were performed to assess the construct validity. The study used descriptive statistics and performed interrater reliability and correlation analyses.

Results: The six groups of evaluators were mostly consistent in their assessments, which clustered into two distinct factors: individual performance (Factor 1) and team/service performance (Factor 2). Factor 1 comprised evaluations from peers, teaching assistants, primary instructors, and group instructors, emphasizing individual students' attendance, participation, and contribution throughout the course. Factor 2 comprised evaluations from service institutions and final report evaluators, focusing on group-level service outcomes and teamwork effectiveness. These two factors explained a cumulative variance of 77.94%. The study identified 15 correlation coefficients: 8 were significantly positive—indicating agreement within or across factors; 2 were significantly negative—highlighting potential divergences in perspective; and 5 were nonsignificant. The relationship between evaluator and student significantly affected assessment outcomes. For instance, peer assessments were the most variable due to subjective influences such as interpersonal dynamics and collaboration history, whereas group instructor assessments showed the least variability, possibly due to a more outcome-focused evaluation approach.

Conclusion: Assessments by different types of evaluators are relatively consistent, and the evaluator–student relationship influences assessment outcomes.

Keywords: project-based service-learning, medical education, multistakeholder assessment, assessment methods, interrater reliability

Introduction

Service-learning is an educational approach that involves both classroom learning and community service and enables students to learn while serving and to enhance their practical problem-solving skills.[1] In medical education, students participating in community service can gain a deeper understanding of societal needs, strengthen their self-directed learning abilities, and develop a sense of social responsibility.[2–4] Service-learning also provides students with early exposure to clinical practice, allowing them to grow through interactions with patients, families, and health-care teams.[1–6]

Project-based service-learning (P$_j$BL-SL) is an instructional approach that integrates the principles of project-based learning (P$_j$BL) with community service. Grounded in John Dewey's educational philosophy of "learning from experience", P$_j$BL-SL emphasizes student autonomy, collaborative problem-solving, and reflective practice to achieve both educational goals and community-oriented outcomes.[1] P$_j$BL-SL instructors and schools collaborate with students to establish learning objectives based on predetermined educational goals and outcomes. Instructors and teaching assistants (TAs) in P$_j$BL-SL courses are responsible for communicating, negotiating, guiding, and facilitating student learning. They also assist students in communicating with institutions receiving the service (hereinafter service institutions).[7–9] P$_j$BL-SL emphasizes student agency, encouraging proactive participation in course learning and the application of acquired knowledge to real-life and work settings, with the goal of enhancing the service-learning outcomes.[1,5,8] P$_j$BL-SL course design includes selecting service projects on the basis of community needs, adjusting course content and skills to be cultivated, promoting mutual learning, providing continual opportunities for reflection, formulating action plans, evaluating the effectiveness of service projects, and celebrating the successful completion of service projects.[10,11] The outcomes of P$_j$BL-SL include increased student motivation, improved communication skills, and enhanced social and emotional learning abilities.[1]

In general, the learning outcomes of service-learning courses are evaluated through self-reflection questionnaires completed by students or by assessing the final service outcomes.[1] Project-based learning course outcomes are examined using qualitative research methodologies, such as interviews with students and teachers and classroom observations; quantitative research methodologies, such as teaching evaluations; and mixed research methodologies, such as using qualitative research to explore influencing factors, followed by quantitative research to verify the relationships between these factors.[12–15] Relying solely on self-reported questionnaires or a single assessment method does not provide a comprehensive understanding of students' learning performance and outcomes in P$_j$BL-SL.

In medical education, student learning is influenced by stakeholders in the educational environment.[16] In service-learning courses, assessments of student learning performance and outcomes by both internal and external stakeholders offer a holistic evaluation of the learning process and results. Internal stakeholders of P$_j$BL-SL courses (ie, students, teachers, and TAs) can evaluate students' learning process, level of engagement, and contributions, and external stakeholders (such as service institutions and final report reviewers) can provide assessments of performance and outcomes in real-world contexts. A multifaceted assessment by both internal and external stakeholders addresses the limitations of a single assessment method and enables accurate measurement of whether students have achieved the objectives of P$_j$BL-SL.

To comprehensively address the contribution of a multifaceted assessment model to student learning outcomes on the basis of multiple stakeholders' perspectives, this study adopted 6 methods to evaluate student learning outcomes in P$_j$BL-SL, namely assessments through group instructors, within-group peers, TAs, service institutions, primary instructors, and final report evaluators. Accordingly, the following 2 research questions were explored:

Research Question 1: What are the differences among the various assessment methods and the consistencies of assessment outcomes generated from these methods?

Research Question 2: Does the relationship between the evaluator and those being evaluated (eg, teacher, TA, peer, service institution) affect assessment outcomes?

This study's contribution lies in breaking through traditional single-assessment models by incorporating perspectives from multiple stakeholders, thereby providing a more holistic and comprehensive way to evaluate medical students' P$_j$BL service-learning. This approach serves as a foundation for fostering student learning and growth and for designing future courses.

Materials and Methods
Course Design
Taiwan's medical education system predominantly follows a six-year undergraduate program that admits students directly upon completion of secondary education. Upon successful completion of the program, graduates are awarded

the degree of Doctor of Medicine (MD). Within this educational framework, service learning has been incorporated into the early stages of medical training to foster students' social responsibility and community engagement.

The 2024 (spring semester) service-learning course (required, 1 credit) for first-year students at the School of Medicine at China Medical University, Taiwan, is based on P$_j$BL-SL principles.[1,7–9] The course is taught in 5 stages. In the first stage, 2 lead instructors introduce service-learning and share relevant experiences. In the second stage, teachers and TAs assist students in establishing relationships with service institutions and in drafting a service-learning plan. In the third stage, students formulate a service-learning project proposal. The fourth stage involves the implementation of the service-learning plan and reflection on the service-learning plan. In the fifth stage, students give a presentation to show what they have learned and undergo a comprehensive review. In this study, the students in the course were divided into 11 groups (approximately 12 to 13 students per group), with the service theme focused on promoting health among older adults. Each group included a second-year medical student who had already completed the service-learning course in spring 2023 and who acted as a TA, along with a medical school faculty member serving as the group instructor. Each TA underwent preparatory training aimed at developing the requisite pedagogical competencies and project management skills necessary to effectively support first-year medical students in the planning and execution of their service-learning projects.

Participants

The participants, comprising the stakeholders of the 2024 spring semester service-learning course (a required 1-credit module) at the School of Medicine, China Medical University, included 126 first-year students enrolled in the course, 11 group teaching assistants, 11 group supervisors/teachers, 2 lead instructors, 6 representatives from the service recipient organisations, and 3 evaluators for the final project presentations.

Assessment Methods

This study incorporated various design and validation approaches to ensure content validity, construct validity, and both convergent and discriminant validity of the overall assessment:

Content Validity

In November 2023, the present study invited two service-learning experts (ie, faculty members with over ten years of experience in designing and teaching service-learning courses), two representatives from service institutions (ie, individuals responsible for coordinating service activities within the community organizations that collaborated with the course), one psychometrics expert, and two medical students who had previously served as service-learning teaching assistants to form a relevant population and expert group.[17,18] Three steps were taken to ensure the content validity of the six evaluators.[19]

(1) Multifaceted Assessments: To comprehensively evaluate student learning outcomes in project-based service-learning, six distinct assessment methods were employed, each corresponding to the roles and perspectives of key stakeholders involved in the course. These included within-group peer evaluations, teaching assistant assessments, service institution feedback, and evaluations by primary instructors, group instructors, and final report reviewers. Each group of evaluators assessed students based on observable behaviors and products aligned with their specific interactions with the students. For example, peers, TAs, and instructors assessed individual students' participation attitudes and contributions during group work and course activities. In contrast, service institutions and final report evaluators focused on group-level performance, including service effectiveness, quality of planning and execution, and the coherence and impact of the final deliverables. The assessment purposes, criteria, and rubric components for each method are summarized in Table 1. This integrative approach captures both individual and team dimensions of learning, enhancing the objectivity and comprehensiveness of the evaluation process.

All six evaluator types used a 0–100 scoring scale. While peer assessments, teaching assistants, and instructors (both primary and group) employed shared core indicators—participation attitude and contribution—to evaluate individual-

Table 1 Assessment Purposes and Rubrics for Multimethod Assessment in Medical Project–Based Service-Learning

Assessment Method*	Purpose	Rubrics	Assessment Type
Within-Group Peers	Students assess the performance of their peers within the same group, providing insights into each student's level of engagement and contribution to group collaboration.	1. Participation attitude[%] 2. Contribution[#]	Individual
TAs	TAs provide in-depth observations of students' actual performance, supplementing areas that may be overlooked in assessments by instructors (both primary and group instructors).	1. Participation attitude 2. Contribution	Individual
Primary instructors	The primary instructor assesses students' overall learning attitude and participation throughout the course.	1. Participation attitude 2. Contribution	Individual
Group instructors	Group instructors evaluate students' learning attitude and participation during group collaboration.	1. Participation attitude 2. Contribution	Individual
Service institutions	The service institution provides direct feedback to confirm whether students have effectively met the needs of the service recipients.	1. Team performance 2. Achievement of service objectives 3. Quality of service outcomes 4. Clarity and thoroughness of service activity planning	Group
Final report evaluators	Three service-learning experts evaluate students' final reports on their service-learning outcomes.	1. Topic selection 2. Appropriateness, coherence, creativity, and feasibility 3. Effectiveness of service outcomes 4. Reproducibility 5. Student reflection	Group

Notes: *: Each assessment score ranges from 0 to 100. %: Participation refers to an assessment of the level of commitment and proactiveness displayed by students in team activities. This includes factors such as regular participation in discussions, willingness to take on group tasks, and sharing ideas or making suggestions during team collaboration. The sub-rubrics for evaluating participation cover 3 aspects: Proactiveness (ie, whether the student actively participates and shows interest), attendance and engagement (ie, whether the student participates punctually and is fully engaged), and teamwork (ie, whether the student collaborates well with others and respects others' opinions). #: Contribution refers to the student's actual input to group tasks and outcomes, focusing on the value and specific contributions they bring to the team activities. The sub-rubrics for evaluating contribution cover 3 aspects: professional contribution (ie, whether the student provides professional knowledge or skills that help the group make progress), problem-solving assistance (ie, whether the student actively participates in addressing challenges), and task completion and quality of work (ie, whether the student completes assigned tasks on time and to a high standard).

level performance, service institutions and final report evaluators used role-specific criteria to assess team-level outcomes and report quality. These criteria are detailed in Table 1. All statistical analyses were conducted using raw scores without transformation or aggregation.

(2) Development of Rubrics: The group members collaboratively developed rubrics for the 6 methods to ensure that the assessment criteria aligned with the course learning objectives and expected learning outcomes (see Table 1).

(3) Assessment of Rubrics: Following the development of the rubrics, instructors and TAs who had previously participated in the service-learning course tried out and modified the rubrics.

Construct Validity

This study used confirmatory factor analysis to examine whether the latent factors constructed through the 6 assessment methods aligned with students' learning processes and outcomes in P_jBL-SL.[20] The present study conducted analysis with principal component analysis and adopted the varimax method to enhance the interpretability of the results.

Convergent and Discriminant Validity

This study employed 4 statistical methods to ensure the convergent and discriminant validity of the assessment methods (See Table 2):

(1) Descriptive Statistics: Including mean, standard deviation, and range, to confirm the stability and distribution of the scores.
(2) Interrater Reliability: Using the intraclass correlation coefficient (ICC) to evaluate the consistency of scores across different assessment methods.[21]
(3) Correlation Analysis: Applying Pearson correlation coefficients to examine the linear associations between different assessment methods, ensuring the convergent validity among the methods.[22]
(4) Factor Analysis: Using confirmatory factor analysis to validate discriminant validity, verifying that each assessment method accurately measures distinct constructs.[20]

Institutional Review Board Approval

This study was approved by the Institutional Review Board of China Medical University (CRREC-113-029). All procedures were performed in accordance with the relevant guidelines and regulations. All participants provided informed consent before participation.

Results

Descriptive Statistics

The mean, standard deviation (SD), coefficient of variation (CV), and range for the 6 assessment methods are presented in Table 3. In order to descend SD values, the assessment scores given by within-group peers exhibited the highest variation (8.74), followed by those of TAs (8.05), service institutions (6.82), primary instructors (4.51), group instructors (2.68), and final report evaluators (2.25).

Interrater Reliability

To enhance consistency among evaluators, a calibration process was conducted prior to assessment. Evaluators reviewed and discussed the scoring rubrics to clarify criteria and standardize interpretations. Interrater reliability was subsequently evaluated using the ICC, which yielded a value of 0.626 (p < 0.001), indicating moderate-to-high consistency across different assessment sources.[21]

Table 2 Analytical Framework for Establishing Validity and Reliability in Project-Based Service-Learning Assessment

Statistical Method	Purpose	Variables/Indicators Used	Key Findings Supported
Descriptive Statistics	Assess central tendency and score dispersion	Mean, Standard Deviation (SD), Coefficient of Variation (CV), and Range	Peer and TA assessments most variable; group instructor least variable
Interrater Reliability (ICC)	Evaluate consistency among evaluators	Intraclass Correlation Coefficient (ICC)	Moderate consistency (ICC = 0.626, p <0.05)
Pearson Correlation Analysis	Examine linear relationships between evaluators	15 correlation coefficients	8 positive, 2 negative, 5 nonsignificant correlations
Construct Validity (via CFA)	To assess whether latent constructs representing individual and group performance align with observed learning outcomes, supporting construct validity.	Confirmatory Factor Analysis (CFA) with PCA and varimax rotation	Two-factor structure confirmed: Factor 1 (individual), Factor 2 (team); 77.94% variance explained

Table 3 Descriptive Statistics of Scores Given by Evaluators

Assessment Methods	Mean	SD	Range
Within-group peers	96.32	8.74	70.50
TAs	92.25	8.05	47.00
Service institutions	92.93	6.82	18.00
Primary instructors	89.02	4.51	35.00
Group instructors	93.48	2.68	17.00
Final report evaluators	87.35	2.25	8.03

Correlation Analysis

Analyzing the 6 assessment methods yielded 15 correlation coefficients, of which 8 were significantly positive (all $P < 0.05$). For example, the assessment score provided by within-group peers was significantly and positively correlated with those of primary instructors ($r = 0.929$) and TAs ($r = 0.629$). We observed 2 significant and negative correlations (both $P < 0.05$); for example, the assessment score given by within-group peers was significantly and negatively correlated with that of service institutions ($r = -0.201$). We observed 5 nonsignificant correlations (all $P > 0.05$); for example, the scores given by final report evaluators were nonsignificantly correlated with those of TA ($r = 0.022$), primary instructor ($r = -0.174$), and group instructor ($r = 0.160$; Table 4).

Factor Analysis

The Kaiser–Meyer–Olkin (KMO) measure of sampling adequacy was 0.529, meeting the minimum threshold for factor analysis but indicating limited shared variance among the variables. This suggests that the extracted factor structure should be interpreted with caution. In contrast, Bartlett's test of sphericity was statistically significant ($\chi^2(15) = 513.70$, $p < 0.001$), confirming that the correlations among variables were sufficient to justify factor extraction.

Principal axis factoring extracted two factors with eigenvalues greater than 1, collectively accounting for 77.94% of the total variance. Factor 1 (eigenvalue = 2.70; 45.05% of the variance) included within-group peer assessments, teaching assistant assessments, primary instructor assessments, and group instructor assessments, all of which focused on individual student performance. Factor 2 (eigenvalue = 1.97; 32.89% of the variance) comprised service institution assessments and final report assessments, which emphasized team-based service implementation and outcomes. This two-factor solution reflects the theoretical distinction between individual performance and group-level practice embedded in the course design, thereby enhancing the interpretability and credibility of the findings despite the relatively low KMO value.

Table 4 Correlations Between the Scores Given by Evaluators

Assessment Methods	Within-Group Peer	TA	Primary Instructor	Group Instructor	Service Institution	Final Report Evaluator
Within-group peer	1.000					
TA	0.629*	1.000				
Primary instructor	0.929*	0.548*	1.000			
Group instructor	0.316*	0.534*	0.327*	1.000		
Service institution	−0.201*	−0.025	−0.170	0.458*	1.000	
Final report evaluator	−0.189*	0.022	−0.174	0.160	0.712*	1.000

Note: *: represents a statistically significant correlation at p < 0.05.

Discussion

This study examined the overall consistency of the different assessment methods, the relationships between these methods, and the effect of the relationship between the evaluators and evaluated students on assessment results.

To situate these findings within the context of existing research, our results reinforce and extend previous work on multistakeholder evaluations in medical education. While earlier studies have examined peer or faculty assessments individually, our findings empirically demonstrate how grouping assessors into factors—those emphasizing individual versus group performance—offers a clearer conceptual model. This aligns with Kokotsaki et al, who emphasized the value of structuring project-based learning assessments to capture both student agency and team dynamics.[15] Similarly, Cifrian et al highlighted the pedagogical benefit of integrating different assessment approaches to evaluate students' performance at both individual and collaborative levels in engineering education.[13] By operationalizing a comparable multilevel assessment strategy within a medical education context, our study not only complements these existing frameworks but also advances their applicability to clinical training environments that emphasize both individual accountability and collaborative service outcomes.

Overall Consistency of Different Evaluators

The ICC was 0.620, indicating a moderate level of consistency among the 6 evaluators despite some differences.[23] These score differences arose from the unique rubrics used by each evaluator (Table 1). For example, the rubrics for within-group peers, TAs, group instructors, and primary instructors focused primarily on individual students' participation attitude and contribution. By contrast, the rubrics for service institutions and final report evaluators were more centered on the overall group performance in the service-learning context (Table 1). The results of the confirmatory factor analysis clearly divided the 6 assessment methods into 2 underlying factors.

Scoring Structure of the 6 Assessment Methods

Factor 1: Assessment of Students' Individual Classroom Performance

This factor included within-group peer assessment, TA assessment, group instructor assessment, and primary instructor assessment. The rubrics for these assessment methods were designed on the basis of project-based learning principles, emphasizing individual student performance and contributions throughout the project, focusing specifically on participation attitude and contribution.

Factor 2: Assessment of Students' Service-Learning Performance

This factor included service institution assessment and final report assessment. The rubrics for these assessment methods were developed with an emphasis on the effectiveness of service-learning and team performance, focusing on service activity planning, achievement of service goals, and quality of service outcomes.

Relationships Within and Between Evaluators

Correlation Coefficients Observed in Each Factor

Seven correlation coefficients in the same factor (6 coefficients for Factor 1 and 1 coefficient for Factor 2) exhibited significant and positive correlations ($r = 0.316$ to $.929$), indicating consistency in scores across different assessment methods within any given assessment dimension.

Correlation Coefficients Observed Between the Two Factors

The correlation analysis results between Factor 1 (individual classroom performance) and Factor 2 (service-learning performance) showed that, among the 8 correlation coefficients between the two factors, only group instructor assessment and service institution assessment exhibited a significantly positive correlation ($r = 0.458$). In addition, we observed 2 significantly negative correlations ($r = -0.189$ and -0.201) and 5 nonsignificant correlations.

The significantly positive correlation between assessment methods with respect to the two factors indicated some consistency in the assessment standards across the factors. For example, the positive correlation between group instructor assessment and service institution assessment ($r = 0.458$) suggested that although the assessment standards differed

between the two methods, they yielded consistent student performance scores. This was because both the group instructors and the service institutions evaluated the same group of students. If the entire group of students actively participated in the service-learning activities, they could demonstrate stronger teamwork and planning performance, which in turn led to higher-quality service outcomes as well as higher scores from service institutions.

The lack of significant correlations between student scores given by the different evaluators across the two factors indicated a weak association between cross-factor assessment standards. For example, final report assessment scores were not significantly correlated with those provided by TAs (r = 0.022), primary instructors (r = −0.174), and group instructors (r = 0.160) scores. Given that the TAs, primary instructors, and group instructors all focused on individual student performance in the classroom and group settings, their assessment criteria were consistent. Specifically, the rubrics for these assessments included participation attitude and contribution, both of which pertain to individual student behavior and performance. By contrast, the final report evaluators primarily assessed the overall execution and reflection on the service-learning project, focusing more on students' service outcomes instead of their individual classroom performance.

Between the two factors, we observed significantly negative correlations between students' scores across different evaluators, indicating an inverse relationship of intragroup collaboration with learning-service outcomes and final report performance. For example, scores given by within-group peers had significantly negative correlations with those of service institutions (r = −0.201) and final report evaluators (r = −0.189). Service institutions and final report evaluators focused on overall team performance and achievement of service goal, whereas peers emphasized intragroup collaboration, particularly individual participation and contribution, which was not evaluated by service institutions or final report evaluators. These contrasting perspectives can be partially explained by group dynamics, where perceived imbalances in participation or interpersonal tensions influence peer ratings, and by social desirability bias, which may lead external evaluators to focus on successful outcomes rather than internal group processes. TA observations confirmed that some students received high external ratings for contributing to service outcomes, despite limited visible engagement in group collaboration, resulting in lower peer scores. The collected raw data indicated that some students received high scores from service institutions and final report evaluators but low scores from within-group peers. The TA explained the reason for the aforementioned phenomenon by noting that these students contributed to a successful service outcome but did not participate actively in relevant course activities, leading to lower scores from peers.

The relationship between the evaluators and students may influence the consistency of results across different evaluators and affect the degree of variation within the scores produced from different assessment methods regarding the same factor.

Variations Between Evaluators of the Same Factor

The relationship between evaluators and students can influence the leniency of scores in the same assessment dimension. In descending order of CV values, within-group peer assessment exhibited the highest variation (9.07%), followed by TA assessment (8.73%), primary instructor assessment (5.06%), and group instructor assessment (2.87%). This indicates that peers had the greatest variability, and group instructors had the least variability.

Given Taiwan's predominantly collectivist cultural context, peer evaluations may be influenced by group harmony, relational considerations, and social obligations. Such dynamics might increase the variation in peer-assigned scores, especially in cases where maintaining group cohesion is prioritized over individual performance accuracy. This contrasts with assessment practices in more individualistic cultures, where feedback may be more candid or competitive.

As members of a group, students had direct experience of each other's participation and contribution. Because students observed each other closely and had a personal stake in the group's outcomes, peer assessments may be influenced by subjective factors such as personal impressions, interaction experiences, or collaboration history, leading to the highest variation in peer assessments. Variation in TA assessment was similar to but slightly lower than that of peer assessments because the TAs examined in this study were second-year medical students who completed the service-learning course the previous year, making their perspective on individual student performance similar to that of a peer. The primary instructor's assessment generally focused on students' in-class performance and was based on a more standardized set of criteria. Their ability to evaluate students from a comprehensive perspective and adoption of more

consistent assessment criteria resulted in a lower CV compared with within-group peer and TA assessments. Group instructors, by contrast, concentrated mainly on students' specific performance during group collaboration, leading to even lower score variability.

Regarding the learning-service performance factor, the CV for service institution assessment (2.87%) was higher than that for final report assessment (2.58%). This was attributed to how the outcomes of service-learning exerted a direct effect on the service recipients and their institutions, promoting the service institutions to more deeply perceive the students' service-learning outcomes stronger than did the final report evaluators.

According to the aforesaid study findings and discussion, we compiled the following answers to the research questions:

Answer to Research Question 1: We observed a certain level of consistency among the 6 assessment methods; however, these methods were examined on the basis of two assessment factors, namely students' individual classroom performance (Factor 1) and learning-service performance (Factor 2). The 2 factors not only involved different assessment criteria—individual performance versus team performance—but also provided a more comprehensive assessment of student learning outcomes in P$_j$BL-SL.

Answer to Research Question 2: The relationship between evaluators and students directly affected the differences in assessment results. For example, within-group peer assessments had the highest SD among the 4 assessment methods for individual classroom performance. This phenomenon was due to how students interacted closely with each another, and their assessments were often influenced by subjective factors such as interaction experiences and collaborative relationships, which could lead to greater variability in assessment scores.

Implications

This study employed 6 assessment methods to assess P$_j$BL-SL outcomes. In consideration of budgetary and human resource limitations, implementers of student assessment systems may consider reducing the number of evaluators in actual educational contexts. The correlation coefficient between the scores given by within-group peers and primary instructors reached .929, indicating high similarity between these 2 evaluators. Therefore, we recommend removing primary instructors from future assessment schemes. The ICC for the remaining 5 methods was 0.500 (ICC = 0.452, $P <$ 0.001). For Factor 2, after removing service institution assessment, the ICC increased to 0.590 ($P <$ 0.01); by contrast, after removing final report assessment, the ICC decreased to 0.498 ($P <$ 0.01). Accordingly, we recommend removing service institution assessment to achieve greater consistency among the different types of evaluators. Specifically, the number of evaluator types may be from 6 to 4 to retain within-group peers, TAs, group instructors, and final report evaluators. This streamlined assessment model maintains a comprehensive assessment of both individual and team performance while achieving adequate assessment validity (ICC = 0.590, $P <$ 0.001).

This multistakeholder assessment framework aligns with global educational reforms emphasizing competency-based medical education (CBME), in which multi-source evaluations are essential for capturing complex competencies such as collaboration, accountability, and professionalism. The delineation between individual-level and group-level assessments in this study also echoes the structure of Entrustable Professional Activities (EPAs), highlighting the importance of evaluating both individual accountability and collective performance in authentic learning contexts.

To further strengthen the validity and educational utility of these assessments, medical educators are encouraged to strike a judicious balance between subjective evaluations—which illuminate affective and process-oriented aspects of learning—and objective indicators that ensure fairness and comparability. Multistakeholder triangulation of assessment data offers a promising strategy to overcome the limitations of any single evaluator perspective.

Moreover, this study elucidates the differential stability and reliability of various evaluator roles. Group instructors demonstrated the lowest score variability and moderate positive correlations with external raters, suggesting their potential as dependable anchors in longitudinal evaluation frameworks. While peer evaluations provide granular insights into team dynamics and learner engagement, they may be susceptible to interpersonal bias and should be interpreted in conjunction with more standardized assessments.

The findings of this study offer important implications for the design and implementation of competency-based assessment models in P$_j$BL-SL curricula. By strategically aligning each evaluator's role with clearly defined learning

objectives—such as interprofessional collaboration, reflective practice, and social accountability—educators can ensure that assessment systems not only measure outcomes effectively but also foster the professional development of future healthcare practitioners.

Limitations

This study has several limitations. First, the assessment data were derived from a specific service-learning course context within the Taiwanese medical education system, characterized by a collectivist cultural orientation and centralized curriculum design; thus, the generalizability of findings to other disciplines, teaching models, or cultural and institutional contexts may be limited. Future studies could explore whether similar assessment dynamics emerge in Western or more individualistic settings. Second, no evaluator training was conducted to reduce the influence of evaluator subjectivity on assessment results. These limitations constitute potential topics for future research.

Conclusion

These findings are context-specific to Taiwan's medical education environment and may not be directly transferable to settings with differing cultural norms, institutional policies, or student–teacher power dynamics. Educators seeking to apply similar multistakeholder assessment models in other regions should carefully consider local educational and cultural conditions to ensure contextual relevance and effectiveness.

Education constitutes an ecosystem of stakeholder interactions,[16] and assessments drawn from these multiple perspectives can more comprehensively and precisely capture both individual learning performance and group service achievements in P_jBL-SL courses. Given that P_jBL-SL is fundamentally rooted in collaborative learning, the present study provides a valuable foundation for developing robust assessment models in similar curricula and serves as a critical reference for evaluating student performance in future service-learning programs.

Ethical Approval

This study was approved by the Institutional Review Board of China Medical University (CRREC-113-029). All procedures were performed in accordance with the relevant guidelines and regulations.

Author Contributions

All authors made a significant contribution to the work reported, whether in the conception, study design, execution, acquisition of data, analysis and interpretation, or in all these areas; took part in drafting, revising or critically reviewing the article; gave final approval of the version to be published; have agreed on the journal to which the article has been submitted; and agree to be accountable for all aspects of the work.

Disclosure

The authors report no conflicts of interest in this work.

References

1. Liao SC, Lee MR, Chen YL, Chen HS. Application of project-based service-learning courses in medical education: trials of curriculum designs during the pandemic. *BMC Med Educ.* 2023;23(1):696. doi:10.1186/s12909-023-04671-w

2. Liao SC, Hsu LC, Lung CH. Early patient contact course: 'Be a Friend to Patients'. *Med Educ*. 2008;42(4):1136–1137. doi:10.1111/j.1365-2923.2008.03202.x

3. Gresh A, LaFave S, Thamilselvan V, et al. Service learning in public health nursing education: how COVID-19 accelerated community-academic partnership. *Public Health Nurs*. 2021;38(2):248–257. doi:10.1111/phn.12796

4. Yang YS, Liu PC, Lin YK, Lin CD, Chen DY, Lin BYJ. Medical students' preclinical service-learning experience and its effects on empathy in clinical training. *BMC Med Educ*. 2021;21(1):1–11. doi:10.1186/s12909-021-02739-z

5. Murray J. Student led action for sustainability in higher education: a literature review. *Int J Sust High Educ*. 2018;19(6):1095–1110. doi:10.1108/IJSHE-09-2017-0164

6. Lin L, Shek DT. Serving children and adolescents in need during the covid-19 pandemic: evaluation of service-learning participants with and without face-to-face interaction. *Int J Environ Res Public Health*. 2021;18(4):2114. doi:10.3390/ijerph18042114

7. Barak M. Problem-, project-and design-based learning: their relationship to teaching science, technology and engineering in school. *J Probl Based Learn*. 2020;7(2):94–97. doi:10.24313/jpbl.2020.00227

8. Baran E, Correia AP. Student-led facilitation strategies in online discussions. *Distance Educ*. 2009;30(3):339–361. doi:10.1080/01587910903236510

9. Knoll M. The project method project learning method: its vocational education origin and international development. *J Ind Teach Educ*. 1997;34(3):59–80.

10. Miller A. Tips for combining project-based and service learning. Available from: https://www.edutopia.org/article/tips-combining-project-based-and-service-learning-andrew-miller.2016. Accessed July 27, 2024.

11. Pucher R, Lehner M. Project-based learning in computer science – a review of more than 500 projects. *Soc Behav Sci*. 2011;29:1561–1566. doi:10.1016/j.sbspro.2011.11.398

12. Almulla MA. The effectiveness of the project-based learning (PBL) approach as a way to engage students in learning. *Sage Open*. 2020;10(3):2158244020938702. doi:10.1177/2158244020938702

13. Cifrian E, Andres A, Galan B, Viguri JR. Integration of different assessment approaches: application to a project-based learning engineering course. *Educ Chem Eng*. 2020;31:62–75. doi:10.1016/j.ece.2020.04.006

14. Doppelt Y. Implementation and assessment of project-based learning in a flexible environment. *Int J Technol Des Educ*. 2003;13:255–272. doi:10.1023/A:1026125427344

15. Kokotsaki D, Menzies V, Wiggins A. Project-based learning: a review of the literature. *Improv Sch*. 2016;19(3):267–277. doi:10.1177/1365480216659733

16. Pan GC, Zheng W, Liao SC. Qualitative study of the learning and studying process of resident physicians in China. *BMC Med Educ*. 2022;22(1):460. doi:10.1186/s12909-022-03537-x

17. Burns N, Grove SK. *The Practice of Nursing Research Conduct, Critique, and Utilization*. 2nd ed. Philadelphia: WB Saunders Company; 1993.

18. Yaghmaie F. Content validity and its estimation. *J Med Educ*. 2003;3(1):25–27.

19. Beck CT, Gable RK. Ensuring content validity: an illustration of the process. *J Nurs Meas*. 2001;9(2):201–215. doi:10.1891/1061-3749.9.2.201

20. Harrington D. *Confirmatory Factor Analysis*. Oxford, UK: Oxford University Press; 2009.

21. Liao SC, Hunt EA, Chen W. Comparison between inter-rater reliability and inter-rater agreement in performance assessment. *Ann Acad Med Singap*. 2010;39(8):613. doi:10.47102/annals-acadmedsg.V39N8p613

22. Marasso D, Lupo C, Collura S, Rainoldi A, Brustio PR. Subjective versus objective measure of physical activity: a systematic review and meta-analysis of the convergent validity of the physical activity questionnaire for children (PAQ-C). *Int J Environ Res Public Health*. 2021;18(7):3413. doi:10.3390/ijerph18073413

23. Koo TK, Li MY. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *J Chiropr Med*. 2016;15(2):155–163. doi:10.1016/j.jcm.2016.02.012