

International Journal of Learning, Teaching and Educational Research
 Vol. 25, No. 9, pp. 1003-1026, September 2026
<https://doi.org/10.26803/ijlter.25.9.40>
 Received Jul 29, 2026; Revised Sept 16, 2026; Accepted Sept 17, 2026

Evaluating Lecturer Performance and Psychological Safety in Teacher Professional Education: A Baseline Mixed-Methods Study at Teacher Education in West Java, Indonesia

Asep Herry Hernawan  and Rita Anggorowati* 

Teacher Professional Education Program,
 School of Postgraduate Studies,
 Universitas Pendidikan Indonesia, Bandung, Indonesia

Abstract. During 2022–2023, Indonesia's Teacher Professional Education Program (PPG) underwent a major curriculum revision toward a competency-based model integrated with a Learning Management System, yet empirical evaluations of lecturer performance within this new framework remain limited. This study assessed PPG lecturers' performance as perceived by pre-service teacher students (PPG Calon Guru) across 13 study programs at a research university in West Java, Indonesia, using a convergent mixed-methods design with a questionnaire adapted for the PPG context. The quantitative component comprised a 20-item ordinal survey ($\alpha = .96$) completed voluntarily by 259 students (58.6% response rate from 442 students enrolled in six courses), while the qualitative component involved reflexive thematic analysis of open-ended responses from 252 participants. Lecturers received an average performance score of 95.3%, corresponding to a Very Good rating on the instrument's 3-point response scale, applied uniformly across all 20 items. The Assessment dimension scored highest (96.3%), while Learning Implementation scored comparatively lowest (92.5%, still Very Good), with significant differences observed between courses (Kruskal-Wallis $H(5) = 19.91$, $p = .001$). Thematic analysis identified four themes: inspiration and role-modelling, interactive and contextual learning, planning and structural challenges, and classroom psychological safety. While PPG lecturers demonstrate strong interpersonal and communication skills across all dimensions, targeted support for instructional media and tools and for classroom psychological safety would still address the two areas of comparatively (though not categorically) lower performance. This assessment serves as a baseline for the first of two semesters.

Keywords: lecturer performance; psychological safety; teaching evaluation; teacher education; Indonesia

Citation:
 Hernawan, A. H., &
 Anggorowati, R.
 (2026). Evaluating
 Lecturer Performance
 and Psychological
 Safety in Teacher
 Professional
 Education: A Baseline
 Mixed-Methods Study
 at Teacher Education
 in West Java,
 Indonesia.
*International Journal of
 Learning, Teaching and
 Educational Research*,
 25(9), 1003-1026.
<https://doi.org/10.26803/ijlter.25.9.40>

*Corresponding author: Rita Anggorowati; ritaanggorowati@upi.edu

1. Introduction

Teacher professional education standards are widely acknowledged as a key early-stage factor affecting student achievement in primary and secondary schools (Darling-Hammond, 2017; Tatto et al., 2017). In Indonesia, the *Pendidikan Profesi Guru* (PPG) program is the required postgraduate route to obtaining teacher certification and is designed to ensure that graduates meet the competency requirements set out in Law No. 14 of 2005 on Teachers and Lecturers.

As part of this program, participants are formally referred to as *PPG Calon Guru* (Pre-Service Teacher Professional Education Students), a title created by the Directorate of PPG at the Ministry of Primary and Secondary Education (*Kemendikdasmen*) of the Republic of Indonesia. In this study, the terms "pre-service teacher students" and "*PPG Calon Guru*" are used interchangeably when referring to this group. After a thorough review of its policies, the Ministry of Education, Culture, Research, and Technology launched a restructured PPG curriculum during 2022–2023, making a change from a model which focused on the delivery of content to one that stresses deep learning, differentiated instruction, design thinking, and a growth mindset; these competencies are generally supported in current literature on teacher education (Cochran-Smith & Villegas, 2015; Darling-Hammond et al., 2020).

In the restructured program, the PPG lecturer has taken on a highly central role. They must not only demonstrate the pedagogical skills they teach but also promote transformative professional development and help pre-service teachers develop reflective practice. Nevertheless, despite this central role, Indonesia still has very little systematic assessment of PPG lecturers' performance. Most of the existing research has focused on the outcomes of the PPG program, such as improvements in student competence, developments in TPACK, and students' satisfaction with their field placements, rather than on the performance of lecturers in the classroom as perceived by the students (Sari et al., 2023; Angelina, 2024).

Student Evaluation of Teaching (SET) is one of the most widely used methods worldwide for assessing lecturer effectiveness in higher education (Spooren et al., 2013; Quansah et al., 2024; Leal Filho et al., 2025). Even though SET instruments have come under critical review concerning their validity, possible bias, and connection with grade inflation (Stroebe, 2020; Park & Cho, 2023; Papadogiannis et al., 2024), SET still has merit when it is combined with qualitative student input and is set in the context of the specific program objectives (Sidwell et al., 2025; Sullivan et al., 2024). A mixed-methods approach to SET is strongly advised when the evaluation context is complex, since relying solely on quantitative ratings may obscure important nuances (Creswell & Plano Clark, 2017).

There is increasing concern among those involved in SET research about the impact of classroom psychological safety- that is, the collective belief that it is safe to take interpersonal risks when learning without facing punishment or humiliation (Edmondson, 1999; Arafa & Ali, 2026)- on students' learning

experiences and the results of their evaluations. Psychological safety is particularly important in teacher training; teachers who feel unable to express their opinions or make mistakes in their own university classrooms may be less prepared to create safe learning environments for their future students (Robinson & Held, 2025; McGuire, 2025; Nurgabdeshev et al., 2026).

Against this backdrop, the study addresses the following research questions:

RQ1: *What does the PPG Calon Guru think about lecturers' performance in relation to the five dimensions of teaching quality?*

RQ2: *Is there a significant difference in perceived lecturer performance among the PPG courses and across various student demographic groups?*

RQ3: *How do pre-service teacher students describe their experiences of PPG lecturer performance in their open-ended responses?*

The study empirically assesses PPG lecturers' performance after the reform. It offers suggestions for quality-assurance policy, the design of professional development programs, and future research into SET in higher education in the Global South. This article marks the first phase of a planned two-stage longitudinal evaluation. In keeping with the *Kemendikdasmen* regulation, the PPG *Calon Guru* complete the program over two semesters, enabling the group examined in this study to be followed throughout their professional training. A further measurement is due at the end of Semester 2 (in the fourth week of October 2026), when the same group will be assessed against a different set of courses during their final semester before being certified as licensed teachers.

2. Literature Review

2.1 Student Evaluation of Teaching in Higher Education

Surveys based on the SET method are the primary method for determining students' views on teaching quality at most higher education institutions worldwide (Spooren et al., 2013; Ikram et al., 2026). Meta-analyses show that SET scores moderately correlate with observable teaching behavior and peer evaluations, although effect sizes vary widely by discipline, class size, and instrument design (Uttl et al., 2017). Recent systematic reviews and empirical studies have highlighted ongoing concerns about the SET's validity, including its susceptibility to gender and ethnic bias and the impact of grade expectations on ratings (Quansah et al., 2024; Papadogiannis et al., 2024). However, despite these criticisms, SET remains essential for lecturers' developmental phase when instruments are specifically designed, grounded in real-life conditions, and supported by qualitative data (Sidwell et al., 2025; Sullivan et al., 2024).

Because this study's Learning Implementation dimension includes items on the Learning Management System (LMS) used to deliver assignments, it is also relevant to the wider literature on technology-mediated teacher professional development. Reviews of online and blended learning in teacher education point to inconsistent instructions, uneven digital-literacy support, and administrative burden as recurring implementation barriers, even where the underlying technology performs well.

In professional education settings such as medicine, law, and teacher training, assessment tools must capture the competences specific to professional preparation, not just general pedagogical ones (Samian & Noor, 2012). In PPG programs, this means evaluating lecturers' ability to demonstrate professional teaching behavior, support the development of reflective practitioners, and manage the dual responsibility of mastering academic content and forming professional identity.

Broader reviews of teacher and lecturer performance in higher education reinforce this multidimensional framing. A systematic review of 39 studies on teacher competencies in higher education identified six main competency dimensions and 32 discrete competencies spanning pedagogical, disciplinary, interpersonal, and professional domains. It noted a persistent gap between the competency institutions perceive as important and those assessed in practice (Dervenis et al., 2022). This gap is directly relevant to the present study, which evaluates lecturer performance across five locally derived dimensions using both a closed-ended instrument and an open-ended qualitative strand to reduce the risk of missing important but under-assessed aspects of performance, such as classroom psychological safety.

2.2 Psychological Safety as a Dimension of Teaching Quality

Edmondson first proposed the concept of psychological safety in 1999 in organizational studies, and it has since attracted growing attention in educational research as a key requirement for deep learning and authentic participation (McGuire, 2025). In higher education classrooms, a psychologically safe environment—one in which teachers encourage questions, accept intellectual risk-taking, and avoid public humiliation—is linked to higher levels of student engagement, more creative thinking, and a greater willingness to express genuine uncertainty (Robinson & Held, 2025).

Psychological safety is especially important in teacher training. PPG *calon guru* act as both learners and future professionals, and their pedagogical attitudes are shaped in real time. If a lecturer accidentally creates an atmosphere of fear, he or she not only hinders immediate learning but also implicitly models the kind of authoritarian approach that Indonesia's current curriculum reform aims to remove (Kemendikbud, 2023). This contradicts the *Merdeka Belajar* (Freedom to Learn) philosophy, the main framework of Indonesia's current education reform, which emphasizes student autonomy, critical discourse, and differentiated expression.

Within the five-dimension SET framework used in this study, psychological safety is not treated as a sixth, independently measured dimension; rather, it is conceptualized as a classroom condition that qualitatively cuts across the quantitative Personality and Communication and Classroom Management and Motivation dimensions, particularly the Fostering a Learning Climate item, and is examined through the qualitative strand precisely because a single closed-ended item cannot capture whether students feel safe enough to express disagreement or admit confusion.

2.3 Teacher Professional Education in Indonesia: Policy Context

Government Regulation No. 19 of 2017 reorganized the PPG program in Indonesia, and additional curriculum reforms were carried out during 2022–2023 as part of the Merdeka Belajar policy initiative. The new PPG curriculum arranges the learning around six thematic courses: Philosophy of Education and Values Education; Understanding Learners and Learning; Deep Learning and Basic Assessment; Growth Mindset; Differentiated Learning; and Design Thinking (Kemendikbud, 2023). The courses are delivered in a blended format that includes synchronous sessions led by university lecturers and asynchronous interaction via the national Learning Management System.

The role change required of PPG lecturers in this new structure is considerable. They must move from simply transmitting knowledge to facilitating, from using standardized assessment methods to employing authentic assessment, and from delivering content to mentoring students in developing their professional identities. The current literature does not show how successfully lecturers have made this transition from the perspective of the students they teach, and this study aims to address that gap.

3. Methodology

3.1 Research Design

The study used a convergent mixed-methods design (Creswell & Plano Clark, 2017), involving simultaneous collection of quantitative and qualitative data, separate analysis of these data, and integration of the findings during interpretation. A convergent design was chosen over a sequential alternative (e.g., an explanatory sequential design, in which qualitative interviews would follow and help explain quantitative results) because the single open-ended survey question could be collected from the same respondents at the same time as the quantitative items, at no additional cost to respondents or additional survey administration; a sequential design would have required a separate, later round of data collection and risked a smaller, self-selected qualitative sample. Figure 1 shows the research framework.

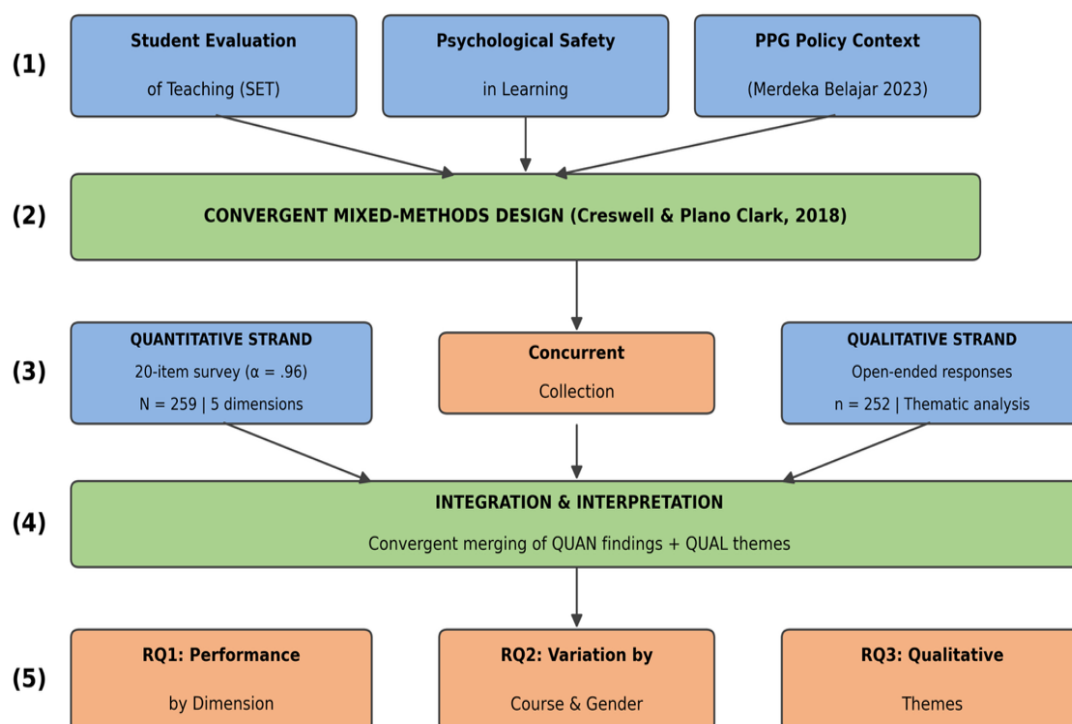


Figure 1: The convergent mixed-methods design and the study's theoretical basis

Note: The five numbered stages are (1) the three theoretical foundations informing the instrument and qualitative question; (2) the convergent mixed-methods design that integrates them; (3) concurrent collection of the quantitative and qualitative strands; (4) integration and interpretation of both strands together; and (5) the three research questions the integrated findings address.

3.2 Participants and Research Site

A total of 442 PPG *Calon Guru* Cohort 1 students were enrolled across all study programs during this period; 259 voluntarily completed the survey via Google Forms (58.6% response rate). Under *Kemendikdasmen* regulation, PPG spans exactly two semesters; this cohort was in Semester 1, with Semester 2 constituting the final semester before certification as a Licensed Teacher. A follow-up measurement of the same cohort, using a parallel instrument covering that semester's courses, is scheduled for the end of Semester 2 (fourth week of October 2026).

These pre-service teacher students, formally designated PPG *Calon Guru* by the Kemendikdasmen Directorate of PPG, were enrolled in the Teacher Professional Education Program, School of Postgraduate Studies, Universitas Pendidikan Indonesia, Bandung, during Semester 1 (Cohort 1) of the 2026 academic year. The coursework began on 2 February 2026. Participants came from 13 study programs across six faculties: Education Studies (Guidance and Counseling, Primary Teacher Education, Early Childhood Teacher Education, Special Education), Social Studies, Science and Technology (Informatics, Software and Game Development, Mechanical Engineering, Automotive Engineering, Culinary Arts), Physical and Health Education, and Arts and Culture. Of the 259 respondents, 204 (78.8%) were female, and 55 (21.2%) were male, consistent with the gender composition of Indonesian teacher-education enrollment.

Evaluations covered six PPG courses: Philosophy of Education and Values Education ($n = 99$), Growth Mindset ($n = 66$), Deep Learning and Basic Assessment ($n = 40$), Understanding Learners and Learning ($n = 33$), Differentiated Learning ($n = 11$), and Design Thinking ($n = 10$). The PPG Program Quality Assurance Team distributed the survey via Google Forms on 4 May 2026; an interim pull on 19 May yielded 165 responses, and after reminders, a final pull on 5 July yielded 259 – a single continuous collection period rather than two cohorts. This is the instrument's first systematic use, establishing a baseline for future PPG quality-assurance cycles.

The PPG Program Quality Assurance Team distributed the survey link to students via WhatsApp and institutional email. The preamble assured participants that responses were confidential and used solely for internal evaluation, that there were no right or wrong answers, and that results would not affect grades or academic standing. Participation was voluntary, with form submission constituting documented informed consent, consistent with accepted ethical standards for non-interventional educational survey research (Creswell & Plano Clark, 2017).

No personally identifiable data beyond study-program affiliation were collected. Because responses are anonymous and cannot be linked across the two measurement points at the individual level, the planned Semester 2 measurement constitutes a repeated cross-sectional assessment of the same cohort and courses rather than an individually linked longitudinal panel; references elsewhere in this manuscript to a "longitudinal evaluation" refer to this repeated-cohort sense.

3.3 Instrument

The 20-item ordinal survey was developed specifically for the PPG context, structured around a five-dimension framework of lecturer competencies drawn from national standards: (1) Learning Planning (3 items; Q1–Q3), (2) Learning Implementation (4 items; Q4–Q6, Q9), (3) Assessment (2 items; Q7–Q8), (4) Classroom Management and Motivation (5 items; Q10–Q14), and (5) Personality and Communication (6 items; Q15–Q20). Each item used a 3-point ordinal response scale (favorable, moderate, unfavorable), scored 3, 2, and 1, respectively, uniformly across all 20 items (see Appendix 1).

The qualitative component comprised a single open-ended question soliciting narrative feedback on lecturer's performance. The instrument developers established content validity through expert review before administration. The Learning Planning items (Q1–Q3) ask students to judge planning as it was enacted in class (whether the course appeared organized and sequenced, and whether materials matched stated objectives), not to audit lecturers' written lesson-planning documents directly, since such documents are not typically shared with students.

We assessed internal consistency using Cronbach's alpha. The full 20-item scale showed excellent reliability ($\alpha = .96$, $n = 259$), with dimension-level alphas ranging from acceptable to excellent: Personality and Communication ($\alpha = .94$), Learning

Implementation ($\alpha = .88$), Learning Planning ($\alpha = .83$), Classroom Management and Motivation ($\alpha = .81$), and Assessment ($\alpha = .76$). The mean inter-item correlation ($r = .53$) indicated adequate item diversity within a coherent construct. A full-scale alpha this high can also reflect item redundancy rather than pure internal consistency, and the two-item Assessment subscale's alpha (.76) is less informative than alpha based on more items, since alpha is a function of both average inter-item correlation and the number of items.

We further examined construct validity through confirmatory factor analysis (CFA) of the proposed five-dimension structure. Given the ordinal item responses, we computed a polychoric correlation matrix. We estimated the model using unweighted least squares (ULS), the recommended approach for ordinal CFA (Forero, Maydeu-Olivares, & Gallardo-Pujol, 2009). CFA was conducted on the full sample ($N = 259$) and showed good fit: $\chi^2(160) = 286.48$, $p < .001$; CFI = .998; TLI = .997; RMSEA = .055; SRMR = .058. Standardized loadings ranged from .86–.96 (Learning Planning), .83–.94 (Learning Implementation), .87–.96 (Assessment), .70–.98 (Classroom Management and Motivation), and .91–.98 (Personality and Communication), all $p < .001$, confirming the proposed structure.

Inter-factor correlations (.83–.99) indicated that the five dimensions, while highly correlated as facets of a general teaching-performance construct, remain distinct enough to justify a five-factor first-order model—independently corroborating the reliability evidence above. The correlation approaching .99 (Learning Implementation with Personality and Communication) is nonetheless high enough to leave open the possibility that these two dimensions are not fully empirically distinguishable in this sample; overall model fit indices establish that the five-factor structure is not contradicted by the data, but do not by themselves establish discriminant validity between the most highly correlated dimension pairs, which would require formal comparison against a more constrained (e.g., four-factor) model.

3.4 Data Analysis

We summarised the quantitative data descriptively at both the item and dimension levels (mean, standard deviation, and percentage of the maximum score). Because the data were ordinal and showed negative skews for all items, we used non-parametric inferential tests: the Kruskal-Wallis H test for differences between courses and the Mann-Whitney U test for pairwise comparisons and gender differences.

We analyzed the qualitative data using Braun and Clarke's (2006, 2022) six-phase reflexive thematic analysis. The 252 open-ended responses were read repeatedly, and the authors developed initial codes inductively based on their semantic content. The authors then gradually organized these codes into candidate themes, examined them across the entire dataset, and ultimately narrowed them to four themes. The two authors discussed and refined theme definitions together throughout the coding process, consistent with reflexive thematic analysis' emphasis on interpretive engagement rather than a search for a single, objectively correct coding structure.

As a supplementary source of triangulation, we also qualitatively reviewed end-of-semester administrative evaluation reports prepared independently by coordinators of eleven participating study areas (program-management documents rather than research instruments) for thematic correspondence with the survey and thematic-analysis findings above; we report the results of this review in Section 5.2. These reports, routinely submitted by study-area coordinators to the PPG Program Quality Assurance Team at the end of Semester 1 (June 2026) as part of standard institutional monitoring, were available for eleven of the thirteen study programs at the time of this review; two programs had not yet submitted their reports.

Coordinators are academic staff responsible for program administration rather than the lecturers being evaluated, and each report follows a standard template with a dedicated section on learning implementation constraints. Use of these existing administrative documents for research triangulation was covered under the same institutional ethical approval described in Section 7.

4. Results

4.1 Overall Lecturer Performance (RQ1)

Table 1 gives the descriptive statistics for all 20 performance indicators. The average score awarded to lecturers, computed as the mean of each item's percentage of its own response-scale maximum, was 95.3 per cent, placing total performance in the Very Good category. Of the 259 respondents, 222 (85.7 per cent) had an overall score in the Very Good range (90 per cent or more of the maximum), 20 (7.7 per cent) gave a score in the Good range, 12 (4.6 per cent) in the Fair range, and 5 (1.9 per cent) in the Poor range.

Table 1: Descriptive Statistics for 20 Lecturer Performance Indicators (N = 259)

Item	Indicator	M	SD	Median	IQR	%	Skew	Category
Q1	Learning Planning (Syllabus & RPS)	2.82	0.47	3	0	94.0%	-2.62	Very Good
Q2	Content Mastery	2.90	0.34	3	0	96.8%	-3.82	Very Good
Q3	Material-Objective Relevance	2.89	0.36	3	0	96.4%	-3.50	Very Good
Q4	Learning Approach and Methods	2.78	0.48	3	0	92.7%	-2.15	Very Good
Q5	Learning Media and Tools	2.73	0.53	3	0	90.9%	-1.79	Very Good
Q6	Learning Resources	2.80	0.46	3	0	93.3%	-2.22	Very Good
Q7	Summative Assessment	2.89	0.35	3	0	96.4%	-3.35	Very Good
Q8	Formative/Process Assessment	2.88	0.38	3	0	96.1%	-3.44	Very Good

Item	Indicator	M	SD	Median	IQR	%	Skew	Category
Q9	Clarity of Assignment Instructions	2.79	0.47	3	0	93.1%	-2.20	Very Good
Q10	Classroom Management	2.88	0.37	3	0	96.1%	-3.32	Very Good
Q11	Teaching Enthusiasm	2.91	0.33	3	0	96.9%	-3.77	Very Good
Q12	Fostering a Learning Climate	2.82	0.49	3	0	94.0%	-2.74	Very Good
Q13	Discipline	2.88	0.35	3	0	96.0%	-2.90	Very Good
Q14	Rule Enforcement	2.90	0.32	3	0	96.8%	-3.43	Very Good
Q15	Fostering Student Character	2.89	0.38	3	0	96.4%	-3.72	Very Good
Q16	Role-Modeling	2.87	0.43	3	0	95.6%	-3.39	Very Good
Q17	Emotional Maturity	2.86	0.46	3	0	95.2%	-3.26	Very Good
Q18	Communication Skills	2.88	0.39	3	0	95.9%	-3.36	Very Good
Q19	Language Use	2.91	0.36	3	0	97.0%	-4.28	Very Good
Q20	Social Interaction Skills	2.89	0.39	3	0	96.3%	-3.70	Very Good
Overall (item mean)		2.86	0.30	3	0	95.3%	—	Very Good

Note. M = mean on a 1–3 scale; SD = standard deviation; median and IQR (interquartile range) are reported alongside the mean given the ordinal, skewed nature of the data; % = percentage of the maximum value (3.00), uniform across all 20 items. Category thresholds: Very Good $\geq 90\%$; Good 75–89%; Fair 60–74%; Poor $< 60\%$. No items fall below the Good threshold. The median is 3 (the maximum), and the IQR is 0 for every item, meaning at least 75% of respondents selected the favorable option on every one of the 20 items; this is a starker illustration of the ceiling effect than the skew statistic alone, and indicates that the mean, while imperfect, is the only one of these three statistics that retains any ability to discriminate between items at this performance level.

Figure 2 shows average scores for all 20 indicators, grouped by dimension. All 20 indicators use the same 3-point response scale and fall in the Very Good range; the widest gap is between Q5 (Learning Media and Tools, 90.9%) and Q19 (Language Use, 97.0%). All 20 items also show pronounced negative skew, consistent with a ceiling effect in students' ratings. The Personality and Communication dimension had consistently high scores.

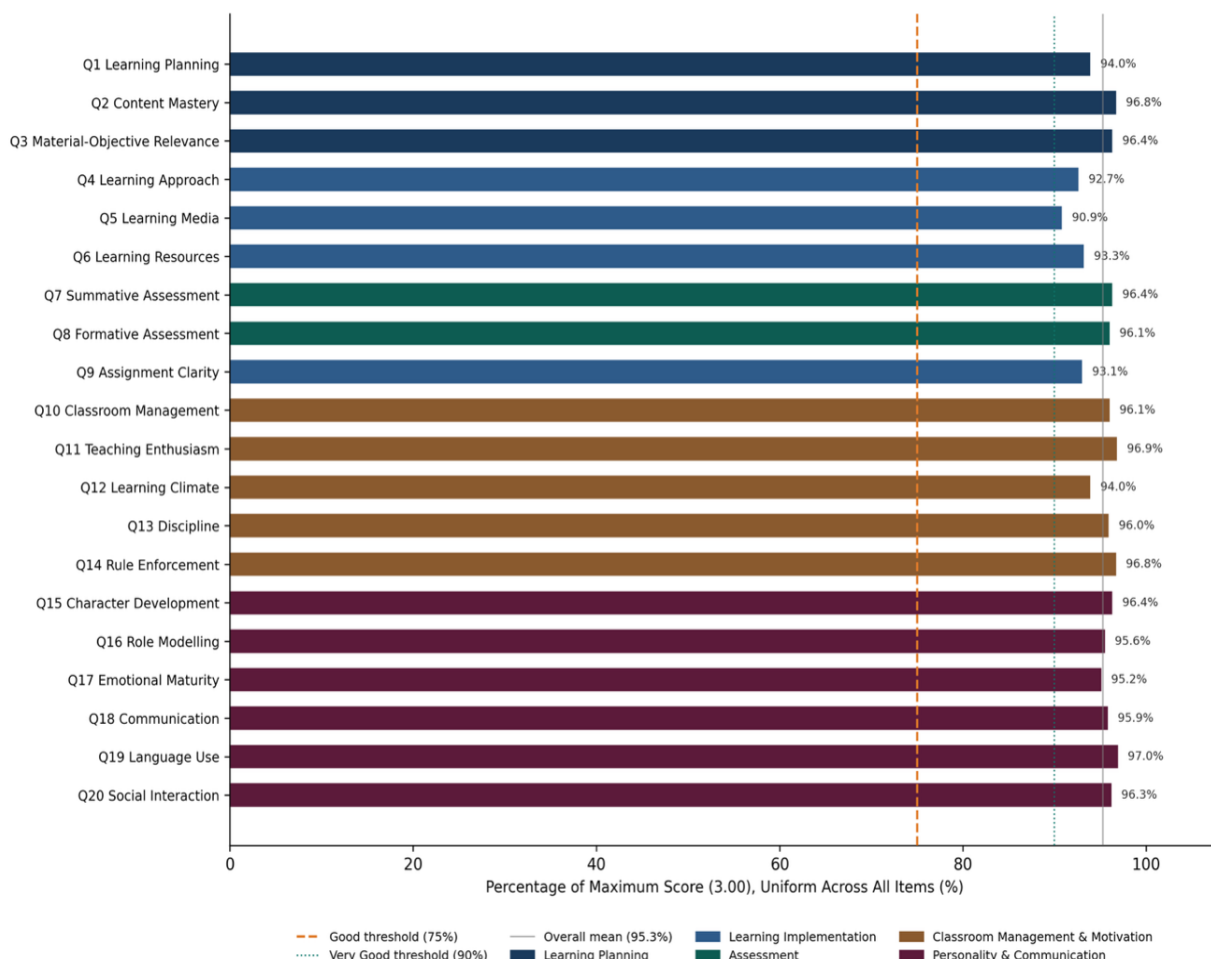


Figure 2: The mean scores for 20 lecturer performance indicators (sample size = 259)

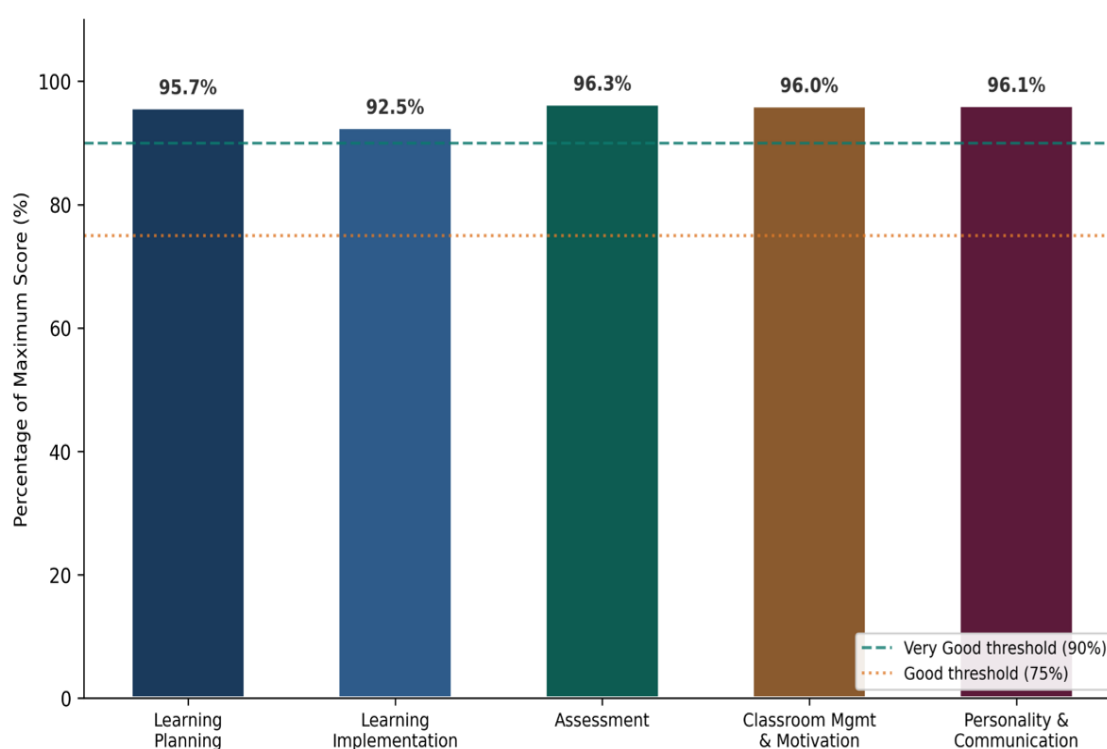
Note: Bar heights and shading reflect each item's percentage of the uniform 3-point response-scale maximum (3.00) applied across all 20 items; no items fall below the good threshold. Dashed vertical lines mark the Good (75%) and Very Good (90%) thresholds; the solid grey line marks the overall mean. Bar colour indicates dimension membership (see legend).

4.2 Performance by Dimension

Table 2 shows the combined performance by dimension, and Figure 3 provides a visual comparison. Every dimension fall in the Very Good range (92.5%-96.3%). Assessment scored highest (96.3%), followed closely by Personality and Communication (96.1%), Classroom Management and Motivation (96.0%), and Learning Planning (95.7%). Learning Implementation had the lowest performance (92.5%, still Very Good), particularly in Question 5 (Learning Media and Tools, 90.9%). Learning Planning itself is not a weakness in this dataset: Q1 (94.0%) and Q3 (96.4%) are both among the higher-rated indicators in the instrument, and Q2 (Content Mastery, 96.8%) is one of the highest-rated indicators overall.

Table 2: Lecturer Performance by Teaching Dimension (N = 259)

Dimension	Items	α	M	%	Category
Learning Planning	Q1–Q3	0.83	2.87	95.7%	Very Good
Learning Implementation	Q4–Q6, Q9	0.88	2.77	92.5%	Very Good
Assessment	Q7–Q8	0.76	2.89	96.3%	Very Good
Classroom Management & Motivation	Q10–Q14	0.81	2.88	96.0%	Very Good
Personality & Communication	Q15–Q20	0.94	2.88	96.1%	Very Good

**Figure 3: Lecturer performance according to teaching dimension (N = 259)**

Note: Dashed lines mark the Very Good (90%) and Good (75%) thresholds. All five dimensions fall in the Very Good range; Learning Implementation is comparatively the lowest (92.5%).

4.3 Inferential Analysis (RQ2)

The Kruskal-Wallis H test indicated a statistically significant difference in overall ratings of lecturer performance across the PPG courses, with $H(5) = 19.91$ and $p = .001$. When the Mann-Whitney U test was applied to the groups pairwise, it was found that the students in the Philosophy of Education and Values Education course gave a significantly lower rating of overall lecturer performance (mean = 2.76, standard deviation = 0.40, $n = 99$) than the students in all the other courses put together (mean = 2.92, standard deviation = 0.20, $n = 160$), the result being $U = 5,761.5$, $p < .001$, which shows that the problem was specific to a particular

combination of lecturer and course and not something that affected the PPG program as a whole. The effect size for the six-course comparison was small to moderate (epsilon-squared = .06).

There was no significant gender difference in the overall ratings of lecturer performance ($U = 6,031.5$, $p = .333$, $r = .053$), which shows that both female students ($M = 2.86$, $SD = 0.31$, $Mdn = 3.00$, $n = 204$) and male students ($M = 2.86$, $SD = 0.26$, $Mdn = 3.00$, $n = 55$) rated the lecturers' performance as equally good. When the analysis was carried out at the level of the individual dimensions, no significant gender differences were found on any of the five teaching dimensions (all $p > .15$). At the item level as well, the pattern was confirmed: none of the 20 separate indicators showed a statistically significant gender difference (all $p > .08$), although Q1 (Learning Planning) came close to but did not reach the significance level, with male students giving it a slightly higher rating than female students ($U = 5,084.0$, $p = .083$).

The consistency of null results across all levels of analysis has substantial meaning: gender does not systematically bias how PPG *Calon Guru* evaluates lecturer performance at this institution. This should not be read as evidence of formal measurement invariance, which would require procedures such as multi-group CFA or differential item functioning analysis; the present finding is limited to the absence of a mean-level gender difference in observed scores. This null result contrasts with parts of the broader SET literature, where gender bias against female instructors has sometimes been reported (Stroebe, 2020); one plausible explanation is that PPG *Calon Guru* students are themselves training to become teachers and may evaluate lecturers primarily on professional competence rather than on characteristics associated with gender bias in general undergraduate SET research (Quansah et al., 2024).

4.4 Qualitative Findings: Thematic Analysis (RQ3)

Out of the 259 respondents, 252 (97.3 per cent) gave qualitative replies. Reflexive thematic analysis identified four themes, as shown by keyword-code frequency in Figure 4. The frequencies reported are counts of responses in which a given theme was identifiable; because a single response could be coded for more than one theme, and because some responses were too brief or generic to code thematically (e.g., "already good," or no substantive content beyond a placeholder), the four theme frequencies in Figure 4 sum to 54.7% of the 252 responses rather than 100%.

The remaining 45.3% comprised responses that were either purely generic/non-substantive or that raised concerns outside the four themes (for example, relevance to field placement, grading transparency, or lecturer attendance) and were not forced into the four-theme structure in order to preserve the interpretive integrity of reflexive thematic analysis.

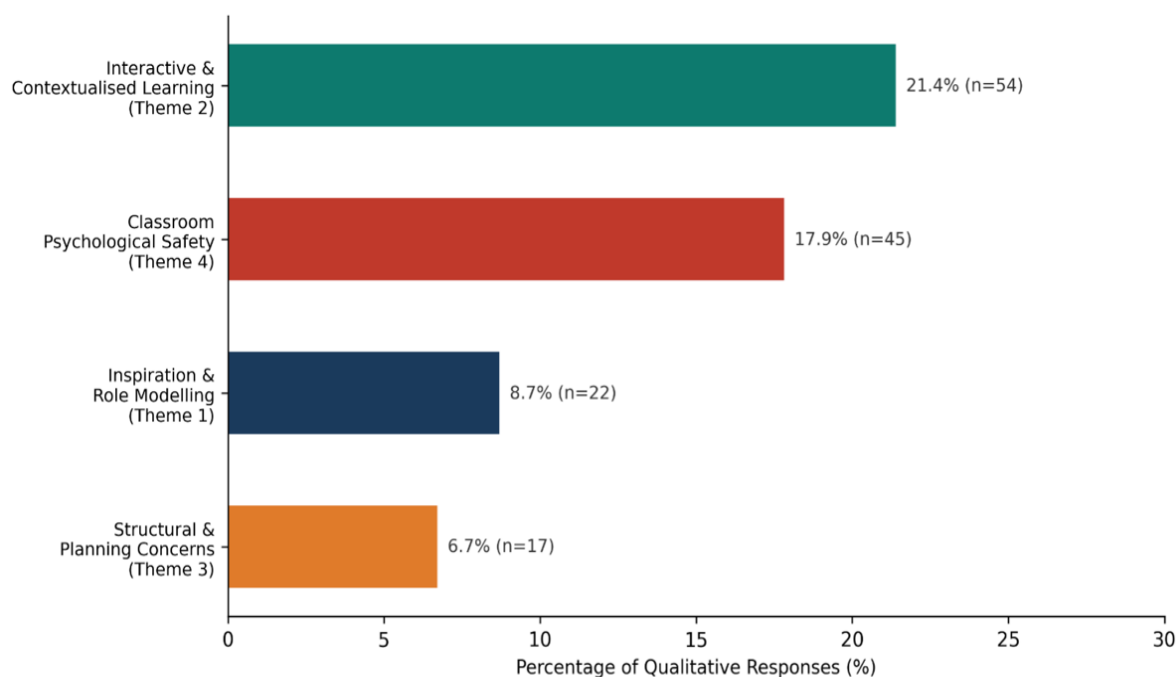


Figure 4: The frequency of qualitative themes obtained from thematic analysis of the open-ended responses (n = 252)

Note: Frequencies represent keyword-coded occurrences; themes were identified reflexively, not by frequency alone. The n = 252 denotes the total number of open-ended responses against which each theme's percentage is calculated, not the sum of the four bars: because a response could be coded for more than one theme, and some responses did not fit any of the four themes, the bars sum to 54.7% of the 252 responses rather than 100%.

4.4.1 Theme 1: The Lecturer as an Inspiring Role Model.

A major theme was the way students saw their lecturers as inspirational people and as professional role models. Often, the lecturers' motivational effect was attributed to their personal character, life experiences, and passion for teaching. *"Truly inspiring. He is a patient, communicative role model who manages to build warm social rapport with students without compromising professionalism."* (PGSD, Growth Mindset course)

"As the lecturer for the Growth Mindset course, [the lecturer K] has already embodied the growth-mindset attitudes and behaviors himself – he has never missed a session, always shows up, and teaches with full commitment." (Software Development, Growth Mindset course)

4.4.2 Theme 2: Interactive and Contextual Learning.

Students consistently preferred learning methods that focused on active discussion, realistic examples, and links to real-world experience. This theme was coded most often (n = 54; 21.4 per cent of responses).

"The lecturer is very receptive to student input, so classroom discussions come alive and open up new perspectives." (Social Studies, Understanding Learners course)

"Classroom learning is highly relevant to what happens in the field, so it was very helpful during teaching practicum." (Physical Education, Philosophy of Education course)

4.4.3 Theme 3: Learning Planning and Structural Issues.

There were worries among the students concerning the discrepancy between the planned program for the course and the way it was actually delivered, the lack of a definite syllabus at the beginning of the semester, the mismatch between the instructions for tasks on the learning management system and those given in class, and the heavy workload.

"It's already quite good, but it would be even better if the course began with an RPS [Semester Learning Plan] so students would know the assignment timeline and what topics would be covered." (PGSD, Philosophy of Education course)

"Content mastery is already fairly deep, but assignments are often inconsistent with LMS instructions. The workload given is often duplicated." (Social Studies, Understanding Learners course)

4.4.4 Theme 4: Psychological Safety in the Classroom.

The fourth theme was expressed most clearly, though not exclusively, in the answers from the Philosophy of Education and Values Education course: students described a classroom atmosphere that was socially tight and inhibited both speaking and intellectual risk-taking. What made these responses stand out was the strong emotional tone and the direct mention of students' well-being.

"I understood the material the lecturer presented, but the communication style and how student opinions were received were too harsh, creating a tense classroom climate where students were afraid to speak up." (Guidance and Counseling, Philosophy of Education course)

"We would like to be able to learn in a comfortable, calm, and enjoyable way – having the courage to speak up and also the courage to be wrong since we are still learning – and at the same time keep our mental health in a good state." (Guidance and Counseling, Philosophy of Education course)

The mention of mental well-being in the last quote is especially important, as it goes beyond teaching effectiveness to raise concerns about students' well-being that warrant institutional attention.

5. Discussion

Following Creswell and Plano Clark's (2017) convergent mixed-methods integration framework, the quantitative and qualitative strands are considered together below in terms of where they converge, complement, and diverge. They converge on personality and communication as clear strengths: the quantitative dimension scores highest, and Theme 1 independently corroborates it qualitatively. They complement one another on Learning Implementation, where a modest quantitative gap (Q5, Learning Media and Tools) is explained qualitatively by Theme 3 and corroborated by independent administrative records. They diverge sharply on classroom climate: the quantitative Learning Climate item scores highly overall, while the qualitative strand surfaces serious

psychological-safety concerns concentrated in one course. This divergence, rather than being a weakness of the design, is treated below as itself a substantive finding about the limits of a purely quantitative SET instrument.

5.1 Strength in PPG Lecturer Performance: Personality as Professional Capital

All five dimensions scored in a narrow, uniformly high band (92.5%-96.3%). Assessment scored numerically highest (96.3%), closely followed by Personality and Communication (96.1%), Classroom Management and Motivation (96.0%), and Learning Planning (95.7%); the differences among these four are marginal. Personality and Communication nonetheless stands out qualitatively: Language Use, Emotional Maturity, and Social Interaction all scored above 95%, showing clearly and consistently that PPG lecturers at this institution have a strong presence in this area.

This result aligns with research suggesting that interpersonal competence is one of the most important factors in motivating students to learn (Samian & Noor, 2012). In professional education, lecturers who demonstrate communicative competence and maintain a disciplined manner serve as excellent examples in teaching practice. Pre-service teacher students do not just observe what lecturers say about teaching; they also note how the lecturers teach (Loughran, 2014; Whitehead & Arslan, 2025). The prominence of Theme 1 (The Lecturer as an Inspiring Role Model) in the qualitative data supports this quantitative result, even though Theme 2 (Interactive and Contextual Learning) was coded more frequently overall.

5.2 Learning Planning and Learning Implementation: Dimension-Level Findings

Learning Planning scored 95.7% (Very Good), with Q1 (Syllabus/RPS delivery, 94.0%) and Q3 (material-objective relevance, 96.4%) both among the higher-rated indicators in the instrument, and Q2 (Content Mastery, 96.8%) among the highest-rated indicators overall. Learning Planning is therefore not a demonstrated weakness in this dataset. The qualitative material in Theme 3 (assignment-instruction mismatches between the LMS and in-class delivery, workload concerns) remains a genuine, separately evidenced concern, distinct from the Learning Planning dimension.

Independent institutional evidence is relevant to the Learning Implementation dimension specifically. Administrative evaluation reports from coordinators of all eleven study areas in this cohort, prepared for program management, not research, and without reference to the survey findings, were reviewed; ten contained an explicit section on learning-implementation constraints, and six of these independently cited LMS-related difficulties (Culinary Arts, Informatics, Software and Game Development, Primary Teacher Education, Special Education, and Physical and Health Education), including incomplete modules, unclear instructions, delayed grading, and access problems for both students and field supervisors. This convergence between student survey data (Q5, Learning Media and Tools, the lowest-scoring item at 90.9%) and independent administrative records suggests that LMS- and media-related implementation friction is a

genuine, if comparatively modest, area for improvement, distinct from the Learning Planning dimension.

5.3 Psychological Safety: A Critical Finding Requiring Attention

It is both important and ironic that psychological safety has become a major issue, since the students who reported this experience were studying in the Guidance and Counseling program, which specifically aimed to promote well-being and foster safe interpersonal relationships. Edmondson (1999) defined psychological safety as a collective belief that taking interpersonal risks will not lead to direct punishment, and the students' responses make this clear: expressions like "afraid to speak up" and "our mental health okay" show that they feel this condition is not present.

Quantitative evidence also supported this: the Philosophy of Education course had the lowest mean score ($M = 2.76$) and the highest standard deviation ($SD = 0.40$), which differed significantly from the other courses ($U = 5,781.5$, $p < .001$). The problem can thus be attributed to a particular combination of lecturer and course rather than to the PPG program. It seems that creating conditions which allow intellectual risk-taking is not only an ethical duty but also a pedagogical advantage, directly linked to Theme 2; in these courses, facilitating active discussion produced the most positive qualitative feedback.

The administrative report prepared by the Guidance and Counseling study area, the same source as the quotations in Theme 4, had no section on learning-implementation constraints (Section 3.4); therefore, this institutional triangulation source cannot confirm or refute the finding regarding psychological safety. This could indicate a structural limitation of formal administrative reports in capturing the classroom climate associated with a particular lecturer or class, compared with technical or logistical constraints that are more easily recorded in official documents. Thus, highlighting the importance of the anonymous student survey as a means of bringing to light aspects of the climate that are not always reflected in top-down administrative reporting.

5.4 Limitations

Several limitations warrant acknowledgment:

First, the data came from a single institution, so transferability to other universities with different lecturer profiles and institutional cultures should be treated with caution.

Second, the cross-sectional design precludes causal inference.

Third, the tab observed across all items suggests a ceiling effect that may limit the instrument's ability to discriminate at higher performance levels.

Fourth, although the 58.6% response rate (259 of 442 students) is acceptable for voluntary online surveys in higher education (Baruch & Holtom, 2008; Al-Abdullatif, 2026), the 41.1% non-response rate raises the possibility of non-response bias, since non-participating students may hold systematically different views.

Fifth, the survey covered only six courses with assigned study-area lecturers; nationally mandated courses delivered independently via the LMS without an assigned lecturer, such as Mental Health Literacy, a prerequisite for the UKPPPG certification exam, fall outside its scope. Administrative reports from the Guidance and Counseling coordinator indicate that this lack of study-area ownership contributed to one student failing to complete the Mental Health Literacy final exam before the issue was resolved in a cross-study-area forum; governance of such independently delivered courses warrants separate evaluation beyond this study.

Sixth, this is the instrument's first application within the institution's PPG quality-assurance process, so read the findings as an initial benchmark rather than a stable assessment. A follow-up measurement is planned for the end of Semester 2 (October 2026), surveying the same cohort across a different set of courses; consistent with the *Kemendikdasmen* regulation limiting PPG to two semesters, this will yield a repeated cross-sectional dataset for the same cohort across two time points, allowing comparison of aggregate perceived lecturer performance and evaluating the impact of quality-improvement interventions introduced in response to this baseline, though not individual-level change tracking.

Seventh, the three-item Learning Planning subscale asks students to rate whether instruction appeared planned and whether materials matched objectives, but does not probe specific planning indicators (e.g., clarity of learning outcomes, alignment between outcomes and content) that students may be poorly positioned to judge directly, since formal lesson-planning documents are not typically shared with them; the dimension is best interpreted as students' perception of planning as enacted in class rather than a direct audit of lecturers' planning documents.

Eighth, the pairwise comparison between the Philosophy of Education and Values Education course and all other courses combined, following a significant six-group Kruskal-Wallis test, was not a pre-registered omnibus post hoc procedure and was motivated by the pattern observed in the data; together with the substantial imbalance in course sizes (from $n = 10$ to $n = 99$), this comparison should be treated as descriptive rather than confirmatory, and the smallest course groups in particular warrant cautious interpretation.

Ninth, because each course in this dataset was taught by a single lecturer, the design cannot statistically separate lecturer-specific effects from course-specific effects; the lower ratings observed in the Philosophy of Education and Values Education course could reflect the lecturer's communication style, the normative/philosophical nature of the course content, or some combination of the two, and the qualitative material (Theme 4) points more toward the former, but this cannot be established definitively from the present design and remains an open question for future research using a design in which lecturers teach multiple courses or courses are taught by multiple lecturers.

Tenth, the planned Semester 2 measurement will cover a different set of courses, with different credit loads, difficulty levels, and in part different lecturers, and will survey a cohort with a semester's additional experience and maturity; any difference observed between the two measurements should therefore be interpreted with these contextual factors in mind, rather than as evidence of a change in lecturer performance alone.

Eleventh, psychological safety itself was not measured using a validated quantitative scale (e.g., Edmondson's original instrument); it is examined here only as a qualitative theme triangulated against one quantitative item (Learning Climate) and course-level comparisons, and a dedicated psychological-safety scale would allow this finding to be assessed more directly in future data collection.

6. Conclusion

This study provides the first empirical evaluation of PPG lecturer performance within Indonesia's post-reform teacher professional education program, drawing on survey responses from 259 students across 13 study programs together with a thematic analysis of 252 narrative responses. Overall, PPG lecturers demonstrated strong performance (95.3% of the maximum possible score across items), including in personality, communication, and motivation and every dimension falls in the Very Good range; Learning Implementation (92.5%) is comparatively the lowest, largely reflecting variability in instructional media and tools (Q5, 90.9%) rather than a broader planning deficit.

A notable and unexpected finding was the emergence of psychological-safety concerns, particularly within the Guidance and Counseling class, where a comparatively assertive lecturer communication style appeared to foster a learning climate in which students felt unable to express their views freely; this has implications not only for learning effectiveness but also for the professional dispositions being cultivated in pre-service teachers who will themselves go on to teach. Concrete next steps for the study area concerned could include structured peer observation of the course paired with feedback focused specifically on the Fostering a Learning Climate indicator, and optional communication-skills coaching for the lecturer involved, with a follow-up qualitative check in Semester 2 to assess whether students' sense of safety to speak up has changed.

As a baseline study within an ongoing institutional quality-assurance cycle, future research should: (a) conduct the planned Semester 2 follow-up assessment in October 2026, covering the different courses offered in the cohort's final semester, to establish a repeated cross sectional dataset for the same cohort and evaluate both performance changes and the impact of quality-improvement measures introduced in response to this baseline; (b) replicate the CFA with the incoming Semester 2 cohort to examine the temporal stability of the instrument's five-dimensional structure; (c) examine variation in PPG lecturer performance across institutions to assess the generalizability of these findings; and (d) investigate the relationship between perceived lecturer quality and pre-service teachers' teaching-competency outcomes during field practicum.

7. Conflict of Interest

The authors declare no conflict of interest. This study was conducted while they were serving as members of the academic staff at Universitas Pendidikan Indonesia. They did not obtain any external funding for it. The host institution was not involved in any aspect of the study, including the design, data collection, analysis, interpretation, or the decision to publish.

8. Acknowledgments

The authors employed Claude (provided by Anthropic, version claude-sonnet-4-6, July 2026) to help with editing the English text, focusing on enhancing its expressive clarity, ensuring grammatical consistency, and adjusting the academic tone. The same AI tool was also used to aid in creating the data-visualization figures (Figures 1–4) based on the main dataset. All the statistical analyses were carried out independently by the authors and can be fully reproduced using the original dataset.

The references were compiled and formatted by the authors themselves. The AI tool played no role in research design, data collection, statistical analysis, data interpretation, thematic coding, or in reaching any conclusions, all of which remain the sole intellectual responsibility of the authors. The use of AI assistance has been disclosed transparently, in line with the target journal's authorship and AI disclosure guidelines, as well as the Committee on Publication Ethics (COPE) recommendations on the use of AI in scholarly publishing (COPE Council, 2023).

9. Ethical Approval

The study was approved by the Research Ethics Commission at Universitas Pendidikan Indonesia, and in keeping with this approval, the survey was carried out with documented informed consent and assurances to all participants regarding confidentiality and the absence of any consequences.

10. Author Contributions

The first author was responsible for conceptualization, instrument design for the quantitative part, supervision, and review and editing. The second author designed the instrument for the qualitative part, conducted data collection, performed formal analysis, carried out data visualization, prepared the original draft, and served as the corresponding author.

11. Data Availability

The anonymized dataset supporting this study's findings is available from the corresponding author upon request, provided participant confidentiality is maintained.

12. References

- Al-Abdullatif, F.A. (2026). No silver bullet for online survey response propensity: evidence from Saudi faculty and graduate students. *Frontiers in Education*, 11, Article 1752542. <https://doi.org/10.3389/feduc.2026.1752542>
- Angelina, P. (2024). Evaluating English pre-service teacher professional education program implementation. *LLT Journal: A Journal on Language and Language Teaching*, 27(2), 1078–1089. <https://doi.org/10.24071/llt.v27i2.7172>
- Arafa, S.M., & Ali, S.I. (2026) “Please don’t expose me”: students’ perceptions of psychological safety during peer feedback on presentations. *Frontiers in Psychology*, 17, Article 1743621. <https://doi.org/10.3389/fpsyg.2026.1743621>
- Baruch, Y., & Holtom, B. C. (2008). Survey response rate levels and trends in organizational research. *Human Relations*, 61(8), 1139–1160. <https://doi.org/10.1177/0018726708094863>
- Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative Research in Psychology*, 3(2), 77–101. <https://doi.org/10.1191/1478088706qp063oa>
- Braun, V., & Clarke, V. (2022). *Thematic analysis: A practical guide*. SAGE Publications.
- Cochran-Smith, M., & Villegas, A. M. (2015). Framing teacher preparation research: An overview of the field, part 1. *Journal of Teacher Education*, 66(1), 7–20. <https://doi.org/10.1177/0022487114549072>
- COPE Council. (2023). *COPE position: Authorship and AI*. Committee on Publication Ethics. <https://doi.org/10.24318/cCVRZBms>
- Creswell, J. W., & Plano Clark, V. L. (2017). *Designing and conducting mixed methods research (3rd ed.)*. SAGE Publications.
- Darling-Hammond, L. (2017). Teacher education around the world: What can we learn from international practice? *European Journal of Teacher Education*, 40(3), 291–309. <https://doi.org/10.1080/02619768.2017.1315399>
- Darling-Hammond, L., Flook, L., Cook-Harvey, C., Barron, B., & Osher, D. (2020). Implications for educational practice of the science of learning and development. *Applied Developmental Science*, 24(2), 97–140. <https://doi.org/10.1080/10888691.2018.1537791>
- Dervenis, C., Fitsilis, P., & Iatrellis, O. (2022). A review of research on teacher competencies in higher education. *Quality Assurance in Education*, 30(2), 199–220. <https://doi.org/10.1108/QAE-08-2021-0126>
- Edmondson, A. C. (1999). Psychological safety and learning behavior in work teams. *Administrative Science Quarterly*, 44(2), 350–383. <https://doi.org/10.2307/2666999>
- Forero, C. G., Maydeu-Olivares, A., & Gallardo-Pujol, D. (2009). Factor analysis with ordinal indicators: A Monte Carlo study comparing DWLS and ULS estimation. *Structural Equation Modeling*, 16(4), 625–641. <https://doi.org/10.1080/10705510903203573>
- Ikram, M., Saleem, S. M. U., & Mehmood, W. (2026). The impact of education quality indicators on students’ satisfaction. *Journal of Applied Research in Higher Education*, 18(1), 326–341. <https://doi.org/10.1108/JARHE-03-2024-0134>
- Kemendikbud. (2023). *Panduan Program Pendidikan Profesi Guru Prajabatan* [Guidelines for the Pre-Service Teacher Professional Education Program]. Kementerian Pendidikan, Kebudayaan, Riset, dan Teknologi [Ministry of Education, Culture, Research, and Technology].
- Leal Filho, W., Trevisan, L.V., Sivapalan, S., Mazhar, M., Kounani, A., Mbah, M. F., Abubakar, I. R., Matandirotya, N. R., Dinis, M. A. P., Borsari, & Abzug, R. (2025). Assessing the impacts of sustainability teaching at higher education institutions. *Discover Sustainability*, 6, Article 227. <https://doi.org/10.1007/s43621-025-01024-z>
- Loughran, J. (2014). Professionally developing as a teacher educator. *Journal of Teacher Education*, 65(4), 271–283. <https://doi.org/10.1177/0022487114533386>

- McGuire, M. (2025). Feedback, reflection and psychological safety: Rethinking assessment for student well-being in higher education. *Assessment & Evaluation in Higher Education*, 51(3), 532–556. <https://doi.org/10.1080/02602938.2025.2548590>
- Nurgabdeshev, A., Ikram, M. and Sarzhanova, G. (2026) From institutional strategies to relational well-being: How implementation quality shapes psychological safety in schools. *Frontiers in Education*, 11, Article 1799502. <https://doi.org/10.3389/feduc.2026.1799502>
- Papadogiannis, I., Vassilakis, C., Wallace, M., & Katsis, A. (2024). Challenges and trends in student evaluation of teaching: Analysis of SET data from the University of Peloponnese. *Information*, 15(9), 576. <https://doi.org/10.3390/info15090576>
- Park, B., & Cho, J. (2023). How does grade inflation affect student evaluation of teaching? *Assessment & Evaluation in Higher Education*, 48(5), 723–735. <https://doi.org/10.1080/02602938.2022.2126429>
- Quansah, F., Cobbinah, A., Asamoah-Gyimah, K., & Hagan, J. E., Jr. (2024). Validity of student evaluation of teaching in higher education: A systematic review. *Frontiers in Education*, 9, Article 1329734. <https://doi.org/10.3389/feduc.2024.1329734>
- Robinson, H., & Held, F. (2025). Psychological safety in online interdisciplinary student teams: What teachers can do to promote an effective climate for knowledge sharing, collaboration and problem-solving. *Active Learning in Higher Education*, 26(3), 445–462. <https://doi.org/10.1177/14697874241275346>
- Samian, Y., & Noor, N. M. (2012). Students' perception of good lecturer is based on lecturer performance assessment. *Procedia - Social and Behavioral Sciences*, 56, 783–790. <https://doi.org/10.1016/j.sbspro.2012.09.716>
- Sari, W. R., Komarudin, K., Badrujaman, A., & Tola, B. (2023). Systematic review of the impact evaluation of teacher profession education programmes in Indonesia. *International Journal of Business, Law, and Education*, 4(2), 555–568. <https://doi.org/10.56442/ijble.v4i2.186>
- Sidwell, D., Lee, D., Zimmerman, P., Bentley, S., & Barton, M. (2025). Teaching faculty experiences with student evaluation of instruction: A mixed-methods study. *Teaching and Learning in Nursing*, 20(1), e276–e284. <https://doi.org/10.1016/j.teln.2024.11.009>
- Spooren, P., Brockx, B., & Mortelmans, D. (2013). On the validity of student evaluation of teaching: The state of the art. *Review of Educational Research*, 83(4), 598–642. <https://doi.org/10.3102/0034654313496870>
- Stroebe, W. (2020). Student evaluations of teaching encourage poor teaching and contributes to grade inflation: A theoretical and empirical analysis. *Basic and Applied Social Psychology*, 42(4), 276–294. <https://doi.org/10.1080/01973533.2020.1756817>
- Sullivan, D., Lakeman, R., Massey, D., Nasrawi, D., Tower, M., & Lee, M. (2024). Student motivations, perceptions and opinions of participating in student evaluation of teaching surveys: A scoping review. *Assessment & Evaluation in Higher Education*, 49(2), 178–189. <https://doi.org/10.1080/02602938.2023.2199486>
- Tatto, M. T., Burn, K., Menter, I., Mutton, T., & Thompson, I. (2017). *Learning to teach in England and the United States: The evolution of policy and practice*. Routledge. <https://doi.org/10.4324/9781315678382>
- Uttl, B., White, C. A., & Gonzalez, D. W. (2017). Meta-analysis of faculty's teaching effectiveness: Student evaluation of teaching ratings and student learning are not related. *Studies in Educational Evaluation*, 54, 22–42. <https://doi.org/10.1016/j.stueduc.2016.08.007>
- Whitehead, G.E.K. & Arslan, S. (2025). Bridging the theory and practice divide in pre-service language teacher education: A focus on teacher educators' expertise from in-service teachers' vantage point. *Social Sciences & Humanities Open*, 12, Article 101656. <https://doi.org/10.1016/j.ssaho.2025.101656>

Appendix 1

The 20-item lecturer performance evaluation instrument was originally administered in Bahasa Indonesia via Google Forms; the English translation below is provided for reference. Each item offered three ordinal response options (favorable, moderate, unfavorable), scored 3, 2, and 1, respectively, uniformly across all 20 items.

Table A1 Twenty-item lecturer performance evaluation instrument (English translation)

Item	Favorable response	Moderate response	Unfavorable response
1. Instructional Planning	Well-planned	Poorly planned	Not planned
2. Content Mastery	Content thoroughly mastered	Content poorly mastered	Content not mastered
3. Material-Objective Relevance	Material and objectives relevant	Material and objectives poorly relevant	Material and objectives not relevant
4. Teaching Approach and Methods	Varied, innovative, creative	Poorly varied, poorly innovative, not creative	Monotonous, not innovative, not creative
5. Learning Media and Tools	Varied, effective, efficient	Poorly varied, poorly effective, poorly efficient	Not varied, not effective, not efficient
6. Learning Resources	Varied, up-to-date, contextual	Poorly varied, poorly effective, poorly efficient	Monotonous, not up-to-date, not contextual
7. Summative Assessment	Aligned with course content and highly objective	Poorly aligned with content and poorly objective	Departs from content and not objective
8. Formative/Process Assessment	Student activity in discussion, observation, and practicum is assessed	...is poorly assessed	...is not assessed
9. Clarity of Assignment Instructions	Clear and structured	Poorly clear and poorly structured	Not clear and not structured
10. Classroom Management	Manages the classroom in an orderly, conducive manner	Poorly able to do so	Not able to do so
11. Teaching Enthusiasm	Highly enthusiastic and motivated to teach	Poorly enthusiastic and motivated	Not enthusiastic or motivated

Item	Favorable response	Moderate response	Unfavorable response
12. Fostering a Learning Climate	Fosters student enthusiasm and a conducive atmosphere	Poorly fosters enthusiasm/atmosphere	Students unenthusiastic; atmosphere not conducive
13. Discipline	High discipline	Poor discipline	No discipline
14. Rule Enforcement	Clear and firm	Poorly clear and firm	Not clear or firm
15. Fostering Student Character	Course fosters student character	Poorly fosters character	Does not foster character
16. Role-Modelling	Good role model	Poor role model	Not a good role model
17. Emotional Maturity	Able to manage emotions	Poorly able to manage emotions	Not able to manage emotions
18. Communication Skills	Communicative	Poorly communicative	Not communicative
19. Language Use	Clear and follows standard spelling/grammar (EYD)	Poorly clear/compliant	Not clear or compliant
20. Social Interaction Skills	Good social interaction	Poor social interaction	No social interaction