

Article

Measuring AI Usage in Geometry Learning Among Pre-Service Mathematics Teachers: Scale Development and Institutional Comparison in Indonesia

Muhammad Ammar Naufal ^{1,*}, Nurfaida Tasni ¹, Andi Syukriani ¹, Anny Sovia ² and Hamzah Upu ¹

¹ Department of Mathematics, Faculty of Mathematics and Natural Sciences, State University of Makassar, Makassar 90224, Indonesia; nurfaida.tasni@unm.ac.id (N.T.); andi.syukriani@unm.ac.id (A.S.); hamzahupu@gmail.com (H.U.)

² Department of Mathematics Education, Faculty of Mathematics and Natural Sciences, State University of Jakarta, Jakarta 13220, Indonesia; annysovia@unj.ac.id

* Correspondence: ammar.naufal@unm.ac.id

Abstract

Pre-service mathematics teachers engage with AI as both geometry learners and future teachers, yet domain-specific instruments for measuring this usage remain scarce. This study developed and validated the AI Usage in Geometry Learning Scale (AIGLS) and compared usage between two Indonesian teacher-education institutions. A comparative survey covered 178 pre-service mathematics teachers at Universitas Negeri Makassar (UNM, $n = 89$) and Universitas Negeri Jakarta (UNJ, $n = 89$). The AIGLS operationalises five dimensions: Conceptual Understanding Support, Problem-Solving Assistance, Geometric Visualisation, Pedagogical Readiness, and Critical Evaluation of AI. A ten-expert panel established content validity ($CVI = 0.976$). Item analysis excluded all five negatively worded items; all analyses therefore used the final 20-item scale, which showed good full-scale (Cronbach's $\alpha = 0.873$) and subscale ($\alpha = 0.591$ – 0.810) reliability. ChatGPT was the dominant tool (96.6%), and UNM reported significantly higher usage intensity than UNJ ($\chi^2(4) = 27.917$, $p < 0.001$). Mann–Whitney U tests with Bonferroni correction confirmed UNM scored higher overall ($U = 4749.5$, $p = 0.022$, $r_{rb} = -0.20$), with Geometric Visualisation showing the largest institutional gap ($r_{rb} = -0.38$, medium effect). The validated AIGLS supports measuring AI usage in geometry teacher preparation and informs AI-literacy policy.

Keywords: geometric thinking; technology-enhanced mathematics education; teacher education; AI literacy; van Hiele theory



Academic Editors: Stavros Kaperonis, Eirini Karakasidou, Fani Gkritseli and Panagiotis A. Tsaknis

Received: 11 June 2026

Revised: 25 July 2026

Accepted: 4 August 2026

Published: 6 August 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

1. Introduction

The integration of Artificial Intelligence (AI) into higher education has fundamentally reconfigured how students interact with academic content. Generative AI platforms have transitioned from experimental adoption to routine learning companions, and a meta-analysis of 22 empirical studies ($N = 5232$) confirmed a moderate positive effect on mathematics learning outcomes ($g = 0.534$) with effects particularly pronounced in geometry contexts (Liu et al., 2026). Within mathematics, geometry is uniquely demanding: geometric reasoning requires simultaneously navigating abstract definitions, deductive proof structures, and dynamic spatial relationships (Battista, 1990; Clements & Battista, 2006; van Hiele, 1986). While generative AI can explain basic geometric concepts effectively,

it struggles with abstract geometric reasoning and frequently produces erroneous proof structures that students accept uncritically (Dilling & Herrmann, 2024; Wardat et al., 2023), underscoring the need for domain-specific investigation of AI usage in geometry learning.

The challenge is especially acute for pre-service mathematics teachers, who engage with AI through a dual lens unavailable to general student populations: as learners developing geometric competency and as future teachers evaluating AI's role in secondary school instruction. As Shulman (1987) established, effective mathematics teaching demands pedagogical content knowledge (PCK) that extends beyond subject mastery; for pre-service teachers, AI usage therefore carries consequences that propagate into future classroom practice. Research confirms that pre-service teachers frequently enter programmes with fragmented geometric knowledge, and that such deficiencies tend to be transmitted to students (Tatto et al., 2012). Tapan Broutin (2024) found that pre-service teachers used AI extensively in lesson planning yet without evaluating output quality, while Drushlyak et al. (2025) demonstrated that AI-based simulation improved critical thinking only when structured guidance was provided. Critically, Bastani et al. (2025) demonstrated that students relying on AI-generated complete solutions performed significantly worse on transfer tasks, and Wijaya et al. (2024) found that higher AI dependency was associated with declining autonomous problem-solving confidence, with risks compounded when the affected individuals are future teachers.

A systematic synthesis reveals four compounding gaps. First, fewer than 12% of AI-in-mathematics studies addressed geometry at university level, and none developed a validated geometry-specific AI usage scale (Nguyen & Pham, 2025). Second, pre-service mathematics teachers are predominantly treated as interchangeable with general learners, despite emerging frameworks conceptualising AI-specific teacher knowledge as qualitatively distinct (Chiu, 2025; Ning et al., 2024). Third, no validated instrument exists to measure AI usage specifically in geometry learning (Turmuzi et al., 2026; Walkington, 2026). Fourth, institution-comparative studies on AI usage in teacher education within developing-country contexts are virtually absent (Celik et al., 2022); Indonesia presents an urgent case given documented digital infrastructure disparities between urban and non-urban universities. Without a domain-specific instrument, the remaining three gaps cannot be addressed rigorously.

This study responds to these gaps directly. First, it introduces and validates a domain-specific, population-targeted instrument for AI usage in geometry learning, anchoring its five dimensions in Bloom's Revised Taxonomy (Anderson & Krathwohl, 2001), Polya's problem-solving model (Polya, 1957), the van Hiele levels (van Hiele, 1986), the TPACK framework (Mishra & Koehler, 2006), and Long and Magerko's (2020) AI Literacy model. Second, it generates the first empirical portrait of AI usage patterns among pre-service mathematics teachers in a geometry learning context. Third, by comparing AI usage across Universitas Negeri Makassar (UNM) and Universitas Negeri Jakarta (UNJ), it provides the first institution-comparative evidence of how regional infrastructure, pedagogical culture, and institutional context shape AI engagement in a specific mathematical domain (Celik et al., 2022; Nguyen & Pham, 2025).

This study is guided by three research questions. RQ1: What are the behavioral and dimensional patterns of AI usage in geometry learning among pre-service mathematics teacher students at UNM and UNJ, encompassing usage frequency, purpose, and scale-level profiles across five theoretically grounded dimensions? RQ2: Is there a statistically significant difference in overall AI usage in geometry learning between UNM and UNJ students? RQ3: On which specific dimensions are institutional differences most pronounced, and what is the practical magnitude of these differences?

2. Literature Review

2.1. AI-Assisted Conceptual Understanding in Geometry

Large language models have expanded access to on-demand conceptual support in mathematics (Kasneci et al., 2023), with a meta-analysis confirming moderate positive effects on mathematics learning outcomes (Li et al., 2025). Within geometry, however, ChatGPT struggles with abstract geometric reasoning and pre-service teachers frequently accept AI-generated proof structures containing logical errors (Dilling & Herrmann, 2024; Wardat et al., 2023). AI tools demonstrably support lower cognitive levels of Bloom's Revised Taxonomy but their capacity for higher-order processes depends heavily on how students engage with outputs (Anderson & Krathwohl, 2001; Walkington, 2026).

2.2. AI in Problem-Solving and Self-Regulated Learning

Intelligent tutoring systems demonstrate consistent achievement gains through adaptive scaffolding (Ma et al., 2014; VanLehn, 2011); however, the risk of over-reliance is well-documented: Bastani et al. (2025) showed experimentally that AI-generated solutions impair transfer learning, and Wijaya et al. (2024) found AI dependency associated with declining problem-solving confidence. Zimmerman's (2000) self-regulated learning framework identifies cyclical goal-setting and reflection as the processes most at risk of being short-circuited by uncritical AI use.

2.3. AI and Geometric Visualisation

Battista (2007) established that spatial reasoning is both a prerequisite and product of meaningful geometric learning, with the van Hiele (1986) model situating visualisation at the entry level. AI-integrated tools, particularly GeoGebra AI and multimodal LLMs, have opened new possibilities for spatial support. Hau et al. (2026) demonstrated that integrating generative AI tools with culturally grounded storytelling significantly enhanced primary students' understanding of circle geometry, with large effect sizes across affective and cognitive dimensions. Yet AI chatbots produce unreliable outputs for abstract three-dimensional relationships (Ergene & Ergene, 2025), underscoring the importance of critical evaluation of AI-generated visualisations. It is important to distinguish two senses of "visualisation" used in this literature: the entry (visual/recognition) level of the van Hiele model, at which figures are apprehended holistically, and the broader Geometric Visualisation construct operationalised in the AIGLS, which additionally encompasses analysis of figural properties and spatial reasoning across multiple van Hiele levels. Empirical evidence at the pre-service teacher level remains limited: Mbatha and Bansilal (2023), for instance, used the van Hiele framework to show that many pre-service mathematics teachers reason about similarity in Euclidean geometry at only the lower visual and descriptive levels, reinforcing the need to examine how AI tools mediate geometric visualisation specifically within this population. Recent syntheses reinforce this priority: Weigand et al. (2025) map how digital and dynamic-geometry environments reshape spatial reasoning at scale, while Naufal et al. (2021) show that embedding metacognitive scaffolds within the van Hiele model significantly raises learners' geometric thinking levels (Awi et al., 2024; Awi et al., 2025; Naufal et al., 2017). At the task-design level, Ulusoy and Girit-Yildiz (2024) further demonstrate that prospective mathematics teachers often under-exploit the cognitive affordances of dynamic-geometry tools, underscoring that access to AI-based visualisation does not by itself guarantee deep geometric engagement.

2.4. AI in Pre-Service Mathematics Teacher Preparation

The TPACK framework (Mishra & Koehler, 2006) remains the dominant lens for technology integration in teaching. The Intelligent-TPACK (I-TPACK) framework (Chiu,

2025) argues that AI's agentic autonomy requires teachers to shift from content deliverers to critical validators of AI-generated information. Empirically, pre-service teachers use AI extensively in lesson planning without evaluating output quality (Tapan Broutin, 2024), and AI-based simulation deepens PCK only when structured evaluation guidance is provided (Drushlyak et al., 2025). Comparative evidence from the Indonesian LPTK context sharpens this concern: Naufal et al. (2025) found significant institution-level differences in the van Hiele geometric thinking levels of generation-Z pre-service mathematics teachers, indicating that geometric competence, and by extension the capacity to critically mediate AI-supported geometry learning, is unevenly distributed across teacher-education institutions.

2.5. Critical Evaluation of AI and AI Literacy

Long and Magerko (2020) defined AI literacy as the capacity to critically evaluate AI outputs, understand AI functioning, and make informed decisions, competencies identified as the most frequently cited gap in AI-in-mathematics research (Turmuzi et al., 2026). For pre-service geometry teachers, this is especially consequential: over-reliance is associated with reduced critical thinking (Abubakar et al., 2025), and the domain is particularly susceptible to AI hallucination including erroneous proofs (Dilling & Herrmann, 2024) and unreliable spatial representations (Ergene & Ergene, 2025).

2.6. Conceptual Framework

The five theoretical streams reviewed above converge into a unified conceptual framework that underpins the AI Usage in Geometry Learning Scale (AIGLS) and the institutional comparison design of this study. As illustrated in Figure 1, each AIGLS dimension is anchored in a distinct theoretical tradition: Bloom's Revised Taxonomy (Anderson & Krathwohl, 2001) grounds the Conceptual Understanding Support (CUS) dimension; Polya's (1957) problem-solving model and Zimmerman's (2000) Self-Regulated Learning framework ground the Problem-Solving Assistance (PSA) dimension; the van Hiele (1986) model and Battista's (1990) spatial reasoning theory ground the Geometric Visualisation (GV) dimension; the TPACK framework (Mishra & Koehler, 2006) and Shulman's (1987) PCK ground the Pedagogical Readiness (PR) dimension; and Facione's (1990) critical thinking framework together with Long and Magerko's (2020) AI Literacy model ground the Critical Evaluation of AI (CEA) dimension. The five dimensions are collectively operationalised as 20 Likert-scale items ($\alpha = 0.873$) forming the AIGLS instrument, which is subsequently applied to compare AI usage patterns between UNM and UNJ through non-parametric inferential statistics.

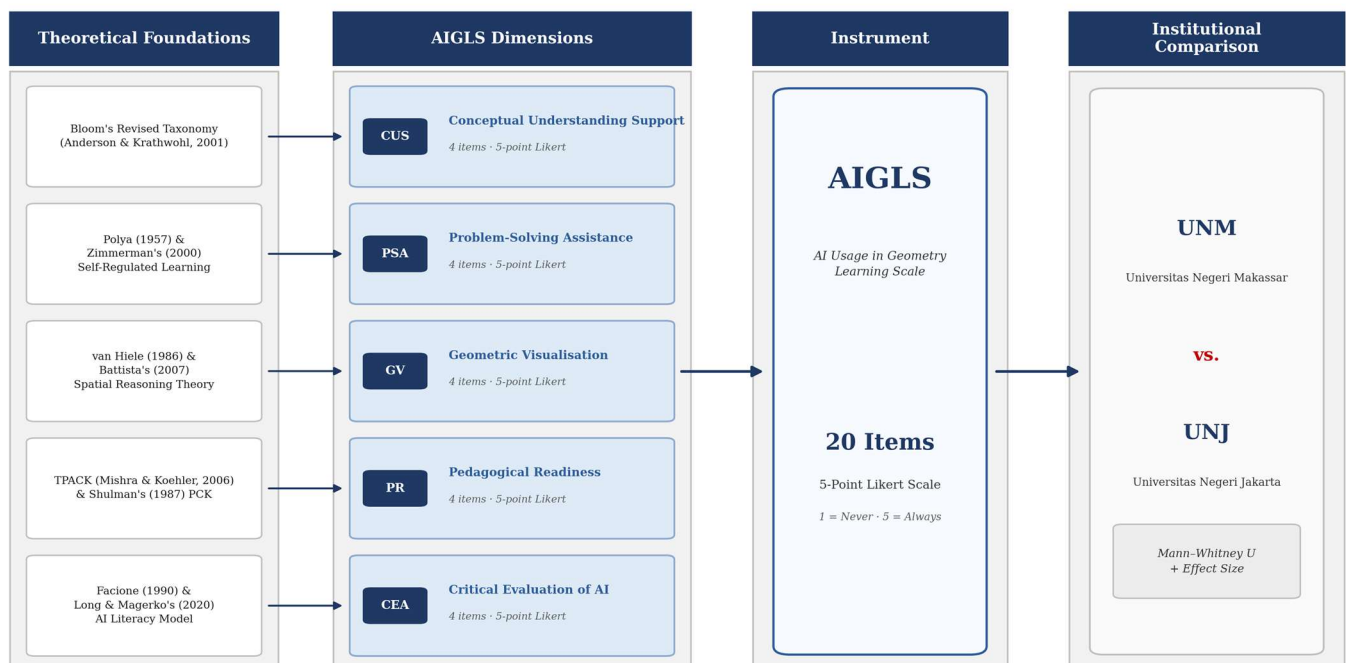


Figure 1. Conceptual framework of the AI Usage in Geometry Learning Scale (AIGLS), integrating Bloom's revised taxonomy (Anderson & Krathwohl, 2001), Polya's problem-solving model (Polya, 1957; Zimmerman, 2000), the van Hiele levels (van Hiele, 1986), Battista's spatial reasoning theory (Battista, 2007), the TPACK framework (Mishra & Koehler, 2006), Shulman's PCK (Shulman, 1987), and the AI literacy model of Long and Magerko (2020) and Facione (1990).

3. Materials and Methods

3.1. Research Design

This study employed a quantitative cross-sectional comparative survey design, appropriate for capturing concurrent snapshots of AI usage behaviours and enabling institutional inference (Bryman, 2016; Creswell & Creswell, 2023). A quantitative approach is further justified by the study's reliance on validated Likert-scale items and the inferential statistical requirements of the comparative research questions.

3.2. Participants and Sampling

The target population comprised active undergraduate students enrolled in the Mathematics Education programme at UNM and UNJ who had completed at least one geometry course and used at least one AI tool for academic purposes. Purposive sampling was employed (Etikan, 2016). Sample size adequacy was verified a priori using G*Power 3.1 (Faul et al., 2009): based on a medium effect size ($d = 0.50$), $\alpha = 0.05$, and power = 0.90, the minimum required sample was 68 participants per group, exceeded by both groups in the final sample. A total of 311 responses were received; after excluding 19 ineligible responses, 292 eligible responses remained. To achieve a balanced comparative design, the UNM sample was randomly subsampled to $n = 89$ without replacement (seed = 42), matching the UNJ sample size, yielding a final analytic sample of $N = 178$ students (UNM: $n = 89$, UNJ: $n = 89$).

3.3. Instrumentation: Development and Validation of the AIGLS

The AI Usage in Geometry Learning Scale (AIGLS) was developed through a systematic multi-stage process grounded in five theoretical frameworks: Bloom's Revised Taxonomy (Anderson & Krathwohl, 2001) for Conceptual Understanding Support (CUS); Polya (1957) and Zimmerman's (2000) SRL for Problem-Solving Assistance (PSA); van Hiele

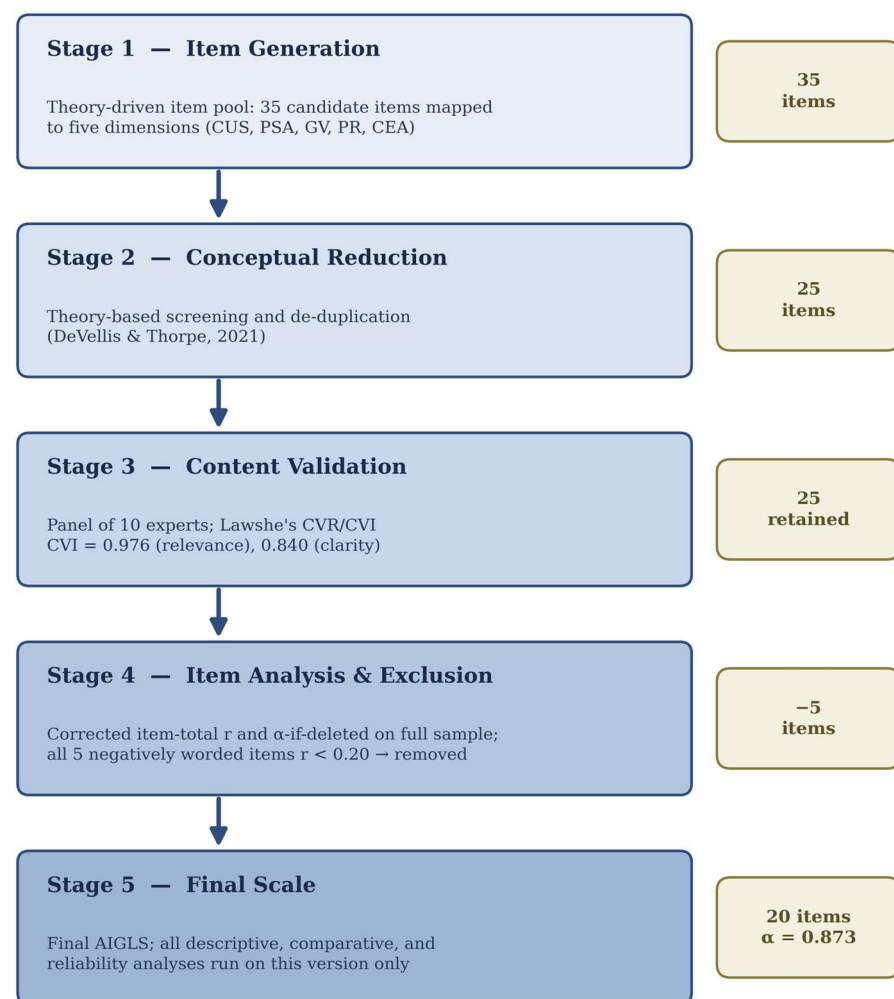
(1986) and Battista (1990) for Geometric Visualisation (GV); TPACK (Mishra & Koehler, 2006) and Shulman's (1987) PCK for Pedagogical Readiness (PR); and Facione (1990) and Long and Magerko (2020) for Critical Evaluation of AI (CEA). An initial 35-item pool was reduced to 25 items through systematic item reduction (DeVellis & Thorpe, 2021). Content validity was established through a panel of 10 expert validators comprising mathematics education lecturers, educational technology specialists, AI experts, and measurement and evaluation experts, using Lawshe's CVR methodology (Lawshe, 1975); all 25 items met the relevance CVR threshold (≥ 0.62) (Wilson et al., 2012), yielding an overall CVI of 0.976 for relevance and 0.840 for language clarity, both exceeding the 0.80 threshold (Polit & Beck, 2006). Nine items were revised based on feedback. Item-level analysis on the main sample revealed that all five negatively worded items exhibited corrected item-total correlations below 0.20 and alpha-if-item-deleted improvements of $\Delta\alpha = +0.12$ to $+0.27$ (Barnette, 2000; Swain et al., 2008); these five items were excluded, yielding the final 20-item AIGLS, with item texts presented in Table 1. Table 1 presents the 20 retained items listed sequentially (No. 1–20), each accompanied by its corresponding dimension, using a 5-point Likert scale (1 = Never, 2 = Rarely, 3 = Sometimes, 4 = Often, 5 = Always). Overall Cronbach's $\alpha = 0.873$ (Excellent per Nunnally (1978)), with subscale alphas reported in Table 2. In summary, the AIGLS was developed through the five-stage pipeline summarised in Figure 2—item-pool generation (35 items) → theory-based reduction to 25 items → expert content validation (Lawshe's CVR/CVI) → item-level analysis and exclusion of the five negatively worded items → final 20-item scale—and all descriptive, comparative, and reliability analyses reported below were computed solely on this final 20-item version.

Table 1. The AI Usage in Geometry Learning Scale (AIGLS).

No.	Item	Dimension
1	I use AI to help clarify geometry concepts that are difficult to understand.	CUS
2	I use AI to explore different representations of geometric shapes (e.g., 2D and 3D).	CUS
3	I use AI to find examples of the application of geometry in everyday life.	CUS
4	I use AI to check my understanding of theorems or geometric formulas.	CUS
5	I use AI to solve geometry problems that I cannot solve on my own.	PSA
6	I use AI to get step-by-step guidance in solving geometry problems.	PSA
7	I use AI to verify the solution steps I have done.	PSA
8	I use AI to find alternative approaches to solving geometry problems.	PSA
9	I use AI to help visualise geometric shapes that are difficult to draw manually.	GV
10	I use AI to generate interactive visualisations or geometric simulations.	GV
11	I use AI to understand geometric transformations through visual representations.	GV
12	I use AI to compare visual representations of different geometric figures.	GV
13	I use AI to help design geometry lesson plans for practicum activities.	PR
14	I use AI to develop geometry teaching materials that are engaging and easy to understand.	PR
15	I use AI to create geometry questions or assessment instruments for students.	PR
16	I use AI to find effective strategies for teaching geometry to students.	PR
17	I think critically about the accuracy of the information provided by AI in geometry.	CEA
18	I cross-check the geometry answers given by AI with textbooks or other trusted sources.	CEA
19	I adjust or correct the output of AI based on my own geometric knowledge.	CEA
20	I am aware of the limitations of AI in solving complex or non-standard geometry problems.	CEA

Table 2. Cronbach’s Alpha and Corrected Item-Total Correlation Ranges for the Initial (25-Item) and Final (20-Item) AIGLS (N = 178).

Dimension	k ₂₅	α ₂₅	k ₂₀	α ₂₀	Δα	r_it Range
CUS—Conceptual Understanding Support	5	0.465	4	0.724	+0.259	0.435–0.567
PSA—Problem-Solving Assistance	5	0.535	4	0.595	+0.060	0.292–0.464
GV—Geometric Visualisation	5	0.687	4	0.810	+0.123	0.575–0.677
PR—Pedagogical Readiness	5	0.632	4	0.779	+0.147	0.553–0.678
CEA—Critical Evaluation of AI	5	0.503	4	0.591	+0.088	0.347–0.403
Total	25	0.814	20	+0.873	+0.059	—

**Figure 2.** Five-stage development and validation pipeline of the AIGLS, from initial item pool to the final 20-item scale. Item reduction followed DeVellis and Thorpe (2021); content validation used Lawshe’s CVR/CVI (Lawshe, 1975).

3.4. Data Collection

Data were collected via an online survey administered through Google Forms over a four-week period. To address common method bias (Podsakoff et al., 2003), respondent anonymity was ensured, items were presented in a fixed non-shuffled order, and negatively worded items were distributed evenly across dimensions. Informed consent was obtained from all participants prior to completing the questionnaire.

3.5. Data Analysis

Reverse coding was applied to all five negatively worded items ($RC = 6 - \text{raw score}$) prior to analysis. The Shapiro–Wilk test indicated non-normal distributions in both groups for all dimensions, necessitating non-parametric inference. Mann–Whitney U tests with Bonferroni correction ($\alpha_{\text{adj}} = 0.05/5 = 0.010$ for dimension-level comparisons; $\alpha = 0.05$ for the total score) were employed for institutional comparisons (Dunn, 1961). Effect sizes were computed as rank-biserial correlations (r_{rb}) with 95% bootstrap confidence intervals (2000 iterations) (Lei & Smith, 2003). Spearman correlations were computed for exploratory inter-dimension analysis (Spearman, 1904; Xiao et al., 2016). All analyses were conducted in Python 3.11 using SciPy 1.11.4.

4. Results

4.1. Instrument Reliability and Item Analysis

Table 2 presents Cronbach’s alpha for both the initial 25-item version and the final 20-item AIGLS. The elimination of all five negatively worded items produced substantial improvements in subscale reliability, with $\Delta\alpha$ ranging from +0.060 (PSA) to +0.259 (CUS). The AIGLS achieved an overall alpha of 0.873. Subscale alphas ranged from 0.591 (CEA) to 0.810 (GV).

4.2. Descriptive Patterns of AI Usage (RQ1)

Table 3 presents mean scores per item (M) and standard deviations (SD) for all five AIGLS dimensions disaggregated by institution, accompanied by 95% confidence intervals. All dimensions fell within the ‘High’ interpretive category, defined as $M = 3.40\text{--}4.19$ on a 5-point scale, for both institutions and the total sample. The highest mean was observed for PSA ($M = 4.174$, $SD = 0.414$), indicating that AI usage is most intensive in problem-solving contexts. The lowest mean was for GV ($M = 3.771$, $SD = 0.617$), reflecting more variable engagement with AI-assisted geometric visualisation. Table 4 presents AI tools reported by the analytic sample; respondents were permitted to select multiple tools, so percentages do not sum to 100%. Six tools were used by at least 10% of respondents. ChatGPT (OpenAI) was used by 96.6% of respondents, followed by Google Gemini (80.9%), Perplexity AI (44.4%), DeepSeek (32.0%), Claude (Anthropic) (24.2%), and Microsoft Copilot (11.2%). Respondents used an average of 3.0 AI tools (range: 1–7), with equivalent multi-tool engagement across both institutions (UNM = 3.0; UNJ = 3.0). The mean score profiles of both institutions across all five dimensions are visualised in Figure 3. These interpretive bands were derived using the standard equal-interval method, in which the response range is divided by the number of response options to yield a class width of $(5 - 1)/5 = 0.80$ (Very Low 1.00–1.79; Low 1.80–2.59; Moderate 2.60–3.39; High 3.40–4.19; Very High 4.20–5.00). Consistent with the reviewer’s guidance, the narrative below foregrounds effect sizes and confidence intervals, treating these descriptive labels as interpretive aids rather than inferential claims.

Beyond tool diversity, the self-reported frequency of AI engagement in geometry learning provides a complementary behavioral indicator. Table 5 presents the frequency distribution disaggregated by institution. Nearly half of all respondents (49.4%) reported using AI 1–2 times per week, while a further 40.4% reported using it 3–4 times per week or nearly every day, indicating a predominantly active engagement profile across the sample. A chi-square test of independence revealed a statistically significant association between institution and usage frequency, $\chi^2(4) = 27.917$, $p < 0.001$, with UNM students reporting substantially higher usage intensity: 22.5% of UNM students reported near-daily use compared to only 2.2% at UNJ. The most frequently cited purposes for AI use were understanding geometric concepts and definitions (79.2%), solving or checking geometry

problems (74.2%), and self-study or examination preparation (73.0%), closely mirroring the CUS, PSA, and GV dimensions of the AIGLS.

Table 3. Descriptive Statistics for AIGLS Dimensions by Institution (N = 178).

Dimension	UNM (<i>n</i> = 89)			UNJ (<i>n</i> = 89)			Total M	Cat.
	M	SD	95% CI	M	SD	95% CI		
CUS	4.211	0.428	[4.122, 4.300]	4.070	0.486	[3.969, 4.171]	4.140	High
PSA	4.191	0.399	[4.108, 4.274]	4.157	0.429	[4.068, 4.246]	4.181	High
GV	3.963	0.582	[3.843, 4.084]	3.579	0.641	[3.445, 3.712]	3.771	High
PR	3.975	0.518	[3.867, 4.082]	3.826	0.567	[3.708, 3.944]	3.900	High
CEA	3.994	0.377	[3.916, 4.073]	4.073	0.480	[3.973, 4.173]	4.034	High
Total	4.067	0.346	[3.995, 4.139]	3.941	0.344	[3.869, 4.013]	4.004	High

Table 4. AI Tools Reported by AIGLS Respondents (N = 178).

AI Tool	UNM (<i>n</i> = 89) <i>n</i> (%)	UNJ (<i>n</i> = 89) <i>n</i> (%)	Total (N = 178) <i>n</i> (%)
ChatGPT (OpenAI)	87 (97.8%)	85 (95.5%)	172 (96.6%)
Google Gemini	69 (77.5%)	75 (84.3%)	144 (80.9%)
Perplexity AI	45 (50.6%)	34 (38.2%)	79 (44.4%)
DeepSeek	25 (28.1%)	32 (36.0%)	57 (32.0%)
Claude (Anthropic)	23 (25.8%)	20 (22.5%)	43 (24.2%)
Microsoft Copilot	10 (11.2%)	10 (11.2%)	20 (11.2%)

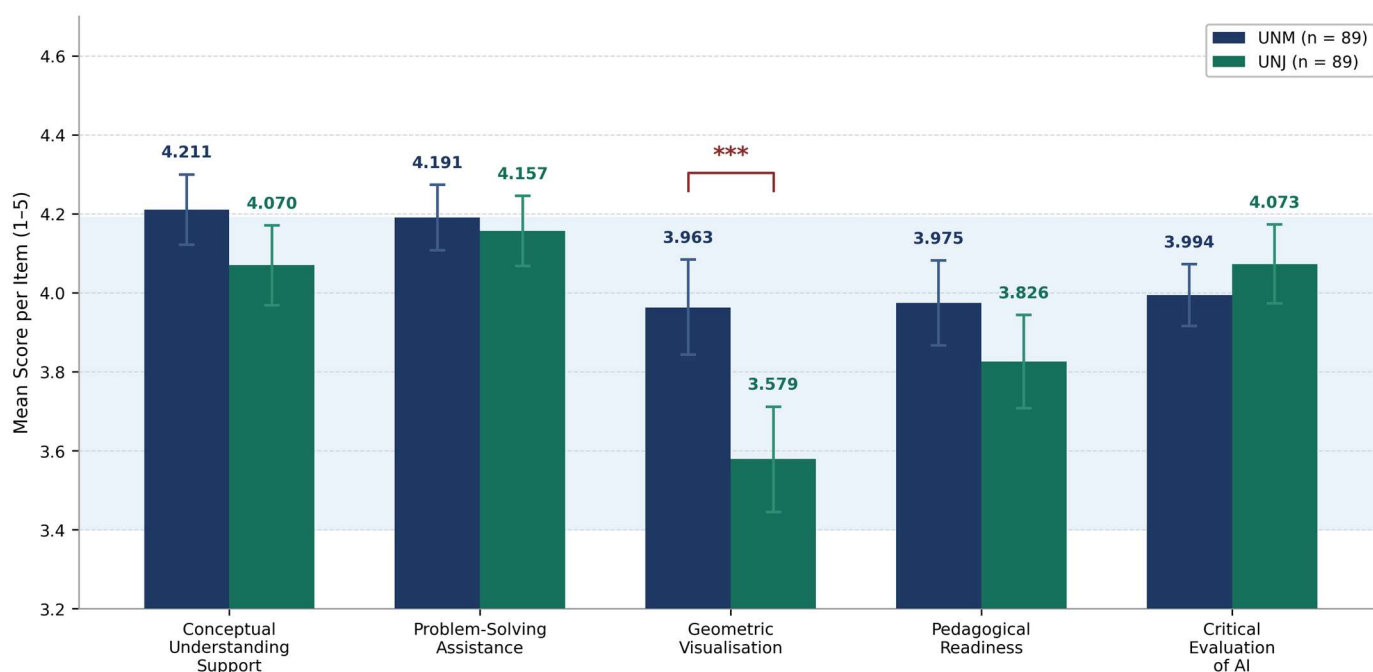


Figure 3. Mean scores per item for each AIGLS dimension by institution (UNM and UNJ) with 95% confidence intervals. The shaded band represents the 'High' interpretive category (3.40–4.19/5.00). *** $p < 0.001$ (Bonferroni-corrected).

Table 5. Frequency of AI Usage in Geometry Learning by Institution (N = 178).

Frequency Category	UNM (n = 89) n (%)	UNJ (n = 89) n (%)	Total (N = 178) n (%)
Never	0 (0.0%)	1 (1.1%)	1 (0.6%)
Rarely (<1×/week)	4 (4.5%)	13 (14.6%)	17 (9.6%)
Sometimes (1–2×/week)	34 (38.2%)	54 (60.7%)	88 (49.4%)
Often (3–4×/week)	31 (34.8%)	19 (21.3%)	50 (28.1%)
Very often (nearly daily)	20 (22.5%)	2 (2.2%)	22 (12.4%)
$\chi^2(4) = 27.917, p < 0.001$			

Together, the frequency and purpose profiles corroborate the dimension-level scale scores and suggest that higher AIGLS means at UNM are partly attributable to greater habitual AI engagement rather than institutional differences in the type of use.

4.3. Overall Institutional Difference in AI Usage (RQ2)

The Shapiro–Wilk test indicated non-normal distributions in both groups for all dimensions, necessitating non-parametric tests. Table 6 reports Mann–Whitney U statistics, two-tailed *p*-values, Bonferroni-corrected significance thresholds ($\alpha_{\text{adj}} = 0.010$ for dimensions; $\alpha = 0.05$ for total), rank-biserial correlations (*r_{rb}*) as effect size estimates, and effect size classifications following Cohen’s benchmarks (Small: $|r| = 0.10$ – 0.29 ; Medium: $|r| = 0.30$ – 0.49). Negative *r_{rb}* values indicate UNM > UNJ. The Mann–Whitney U test revealed a statistically significant overall institutional difference in AIGLS scores between UNM (*M* = 4.067, *SD* = 0.331) and UNJ (*M* = 3.941, *SD* = 0.344) students, *U* = 4749.5, *p* = 0.022, *r_{rb}* = -0.199 , 95% CI [-0.363 , -0.027]. The effect size is classified as small ($|r| = 0.10$ – 0.29) per Cohen (1960), indicating that UNM students report modestly but significantly higher overall AI usage in geometry learning than their UNJ counterparts. The rank-biserial correlation of -0.199 corresponds to a probability of approximately 60% that a randomly selected UNM student will score higher than a randomly selected UNJ student, a non-trivial institutional asymmetry given that both institutions share a national curriculum framework. These findings provide an affirmative answer to RQ2.

Table 6. Mann–Whitney U Test Results: UNM vs. UNJ Comparisons for AIGLS Dimensions (N = 178).

Dimension	UNM M	UNJ M	U	<i>p</i>	Bonf. α	Sig.	<i>r_{rb}</i>	Effect
CUS	4.211	4.070	4615.0	0.051	0.010	ns	-0.165	Small
PSA	4.191	4.157	3985.0	0.943	0.010	ns	-0.006	Neg.
GV	3.963	3.579	5465.5	<0.001	0.010	Sig.	-0.380	Medium
PR	3.975	3.826	4547.5	0.080	0.010	ns	-0.148	Small
CEA	3.994	4.073	3607.5	0.293	0.010	ns	$+0.089$	Neg.
Total	4.067	3.941	4749.5	0.022	—	Sig.	-0.199	Small

Notes: ns = not significant; Sig. = two-tailed significance value.

4.4. Dimension-Level Institutional Differences (RQ3)

To address RQ3, Mann–Whitney U tests were conducted separately for each of the five AIGLS dimensions with Bonferroni correction ($\alpha_{\text{adj}} = 0.010$). Results are presented in Table 6, with effect sizes visualised in Figure 4.

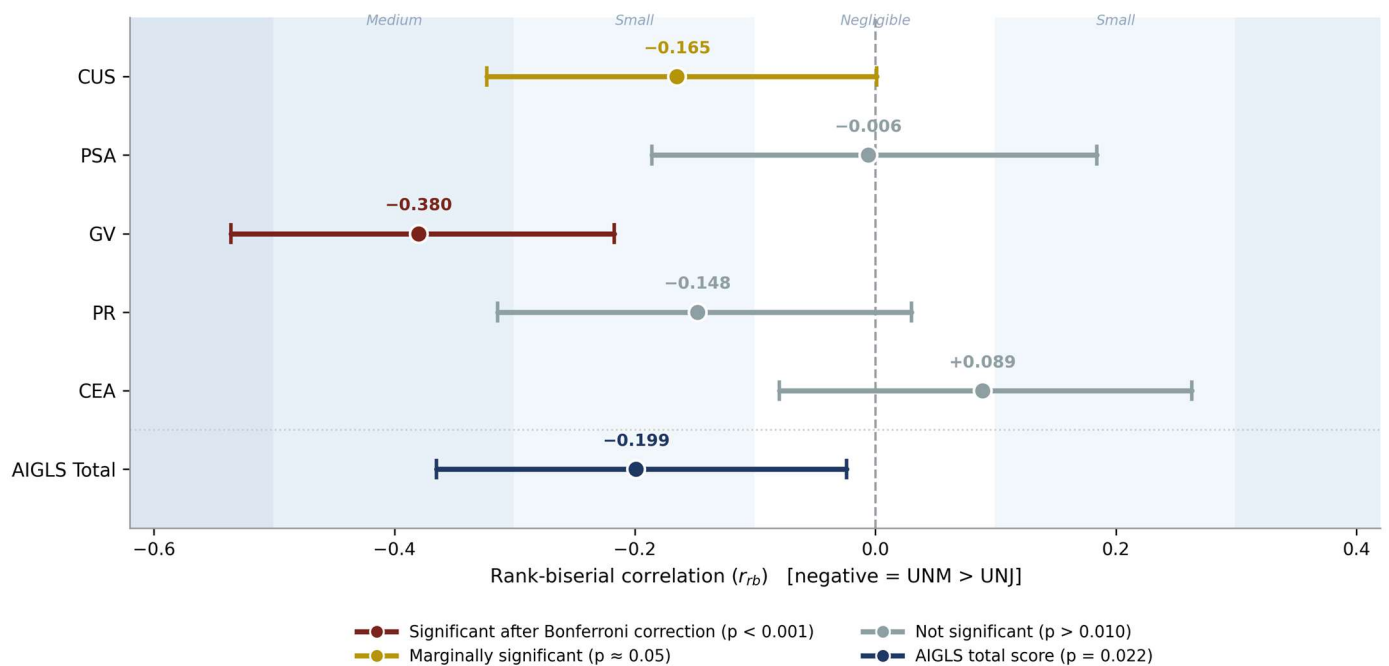


Figure 4. Forest plot of rank-biserial correlations (r_{rb}) with 95% bootstrap confidence intervals for each AIGLS dimension and the total score. Negative values indicate UNM > UNJ. The vertical dashed line marks $r_{rb} = 0$ (no institutional difference); shaded bands denote conventional effect-size magnitudes.

GV was the only dimension to survive Bonferroni correction ($U = 5465.5$, $p < 0.001$, $r_{rb} = -0.380$, 95% CI $[-0.534, -0.219]$, medium effect), with UNM students ($M = 3.963$) substantially outperforming UNJ students ($M = 3.579$) in AI-assisted geometric visualisation. The CUS dimension showed a nominally significant difference ($p = 0.040$) that did not survive Bonferroni correction ($r_{rb} = -0.147$, small effect). PSA, PR, and CEA showed no significant institutional differences (all $p > 0.05$). Notably, CEA showed a reversed direction ($r_{rb} = +0.084$), with UNJ students ($M = 4.073$) numerically outperforming UNM ($M = 3.994$), though this difference was not significant.

4.5. Inter-Dimension Correlations

Table 7 presents the Spearman rank correlation matrix for all five AIGLS dimensions; bold values indicate strong correlations ($\rho \geq 0.50$). All inter-dimension correlations were positive and statistically significant (all $p < 0.001$), ranging from $\rho = 0.235$ (GV–CEA) to $\rho = 0.450$ (CUS–PSA). The consistently weaker associations involving CEA ($\rho = 0.235$ – 0.409) compared with correlations among CUS, PSA, GV, and PR ($\rho = 0.324$ – 0.517) suggest that critical evaluation of AI constitutes a relatively orthogonal competency, consistent with Long and Magerko’s (2020) conceptualisation of AI literacy as a distinct metacognitive construct.

Table 7. Spearman Correlation Matrix for AIGLS Dimensions ($N = 178$).

Dimension	1.	2.	3.	4.	5.	M (SD)
1. CUS	—					4.140 (0.476)
2. PSA	0.450 ***	—				4.174 (0.414)
3. GV	0.442 ***	0.371 ***	—			3.771 (0.617)
4. PR	0.378 ***	0.397 ***	0.517 ***	—		3.900 (0.543)
5. CEA	0.300 ***	0.409 ***	0.235 ***	0.346 ***	—	4.034 (0.429)

Notes: *** $p < 0.001$.

5. Discussion

5.1. Discussion of Descriptive Patterns of AI Usage (RQ1)

All five AIGLS dimensions fell within the High interpretive category ($M = 3.77\text{--}4.17/5.00$), indicating substantive and consistent AI engagement across both institutions. The uniformly high AI engagement across all five dimensions (all $M > 3.80/5.00$) contrasts with earlier research documenting persistent technology adoption barriers in Indonesian higher education (Sutrisno et al., 2025), suggesting rapid normalisation consistent with post-2022 adoption trends (Li et al., 2025). The PSA dimension recorded the highest mean ($M = 4.174$), consistent with Walkington's (2026) finding that problem-solving scaffolding is the most frequently reported AI affordance in mathematics education. However, elevated PSA scores warrant concern: AI-generated complete solutions impair transfer learning (Bastani et al., 2025), and AI dependency is associated with declining problem-solving confidence (Wijaya et al., 2024), with risks compounded for pre-service teachers whose habits will shape future students (Tapan Broutin, 2024). The tool ecosystem data (Table 4) further contextualise these patterns: the near-universal adoption of ChatGPT (OpenAI) (96.6%) alongside moderate uptake of Google Gemini (80.9%) and, to a lesser extent, Perplexity AI (44.4%) and DeepSeek (32.0%) suggests that respondents primarily rely on general-purpose large language models rather than domain-specific mathematical tools. The presence of Claude (Anthropic) (24.2%) and Microsoft Copilot (11.2%) in the tool repertoire, combined with an average of 3.0 tools per respondent, indicates a diversified but LLM-centric AI engagement profile that is consistent across both institutions. The frequency data (Table 5) add a behavioral depth to this picture: the statistically significant institutional difference in usage intensity ($\chi^2(4) = 27.917, p < 0.001$), with 22.5% of UNM students reporting near-daily AI use versus only 2.2% at UNJ, suggests that the higher AIGLS scores at UNM are partly attributable to greater habitual engagement rather than exclusively to qualitative differences in how AI is used.

Two further patterns merit attention. The GV dimension recorded the lowest mean ($M = 3.771$) and the largest standard deviation, reflecting greater inter-individual variability than any other dimension—theoretically explicable because effective AI-assisted visualisation requires familiarity with multimodal tools that have uneven penetration across student populations (Ergene & Ergene, 2025). The CEA dimension, by contrast, shows consistently weaker Spearman correlations with behavioural dimensions ($\rho = 0.235\text{--}0.409$ vs. $\rho = 0.324\text{--}0.517$ among CUS, PSA, GV, and PR), providing empirical support for Long and Magerko's (2020) conceptualisation of AI critical appraisal as a higher-order process independent of usage frequency. Students may use AI extensively while lacking the apparatus to identify proof errors (Dilling & Herrmann, 2024), implying that AI literacy requires explicit scaffolding rather than emerging organically from usage.

5.2. The Psychometric Behaviour of Negatively Worded Items

The systematic underperformance of all five negatively worded items ($r_{it} = -0.07$ to 0.11) is consistent with acquiescence reversal artefacts documented by Barnette (2000) and Swain et al. (2008). Items implicating academic integrity likely trigger social desirability responses (Podsakoff et al., 2003) that suppress item-total correlations, a finding that instructs future development toward behavioural frequency or scenario-based formats. Separately, the relatively modest subscale alphas for PSA (0.595) and CEA (0.591) are interpretable within the constraints identified by Cortina (1993): with only four items per subscale, the mathematical ceiling of alpha is substantially reduced, meaning that these values reflect adequate internal consistency given the item count rather than indicating construct weakness. The corrected item-total correlations for both subscales (0.292–0.464 for PSA; 0.347–0.403 for CEA) confirm that all items contribute meaningfully to their respective constructs.

5.3. Overall Institutional Difference (RQ2)

The statistically significant overall difference ($U = 4749.5$, $p = 0.022$, $r_{rb} = -0.199$, small effect) confirms that institutional context shapes AI usage in geometry learning even when both institutions share a national curriculum framework. The rank-biserial correlation of -0.199 corresponds to a probability of approximately 60% that a randomly selected UNM student will score higher than a randomly selected UNJ student, a meaningful institutional asymmetry with policy implications (Celik et al., 2022). Cohen's (1960) small effect designation should not be interpreted as negligible in the context of cross-institutional comparisons where sampling frames are equivalent, as it reflects genuine population-level differences rather than sampling artefacts.

5.4. Dimension-Level Differences and Institutional Explanations (RQ3)

To address RQ3, Mann–Whitney U tests were conducted for each of the five AIGLS dimensions with Bonferroni correction ($\alpha_{adj} = 0.010$). The results reveal a clear pattern: one dimension accounts for the entirety of the institutional divergence, while the remaining four largely converge, together providing a nuanced portrait of where institutional context matters and where it does not.

The medium-effect GV difference ($U = 5465.5$, $p < 0.001$, $r_{rb} = -0.380$) is the most substantively important finding of the dimension-level analysis. UNM students ($M = 3.963$) substantially outperformed UNJ ($M = 3.579$) in AI-assisted geometric visualisation. Three mechanisms are proposed: differential access to visualisation-specific AI tools (GeoGebra AI, multimodal ChatGPT); differential pedagogical modelling by geometry faculty; and a metropolitan paradox wherein UNJ students may perceive lower utility for AI-assisted visualisation given abundant alternative resources. A fourth contributing factor may be differential access to premium AI tiers: the advanced multimodal capabilities most relevant to geometric visualisation, including image input, interactive spatial reasoning, and 3D model interpretation, are predominantly restricted to paid tiers of ChatGPT (GPT-4o) and Google Gemini Advanced. If UNM students benefit from institutional licensing or greater willingness to subscribe to premium plans, this structural access gap could partially account for the observed GV advantage, independent of individual motivation or pedagogical factors. This extends Battista's (1990) principle that spatial reasoning is more environmentally sensitive than other mathematical competencies into the AI-mediated learning context.

The absence of significant institutional differences on PSA, PR, and CEA (all $p > 0.05$) implies that whatever drives these usage patterns operates through individual-level mechanisms, such as prior AI exposure, general digital literacy, and self-regulation (Zimmerman, 2000), that transcend institutional affiliation; national-level curriculum interventions are therefore the appropriate policy response. The CUS dimension showed a borderline difference ($p = 0.051$, $r_{rb} = -0.165$) that does not survive Bonferroni correction and warrants monitoring in future studies with larger samples. Notably, UNJ students numerically outperformed UNM on CEA ($M = 4.073$ vs. 3.994 , $p = 0.293$), consistent with a metropolitan information environment hypothesis wherein Jakarta students may develop stronger AI scepticism through broader public discourse on AI risks (Abubakar et al., 2025).

5.5. Theoretical Implications

The findings of this study make three contributions to the theoretical literature on AI-integrated mathematics teacher education. First, the CEA dimension's structural dissociation from behavioural dimensions supports I-TPACK (Chiu, 2025) and Long and Magerko's (2020) conceptualisation of AI critical appraisal as a higher-order process independent of usage frequency. Second, the uniformly high AIGLS scores shift the policy

question from whether pre-service teachers adopt AI to whether they engage with it in ways that enhance competency development (Zawacki-Richter et al., 2019). Third, the institution-specific GV gap extends Battista's (1990) principle that spatial reasoning is more environmentally sensitive than other mathematical competencies into the AI-mediated learning domain.

5.6. Practical Implications

The convergence of PSA, PR, and CEA across institutions supports national-level curriculum interventions applicable uniformly across all LPTKs, including integration of AI literacy modules into geometry courses, embedding AI-specific academic integrity training, and formalising pedagogical AI integration within TPACK-based Professional Practice components. At the institution-specific level, the medium-effect GV gap at UNJ ($M = 3.579$ vs. UNM $M = 3.963$) identifies a concrete intervention target: UNJ geometry faculty should receive professional development on AI-assisted visualisation tools, with emphasis on pedagogical modelling of GeoGebra AI and multimodal ChatGPT. Given the strong GV–PR correlation ($\rho = 0.517$), improving GV engagement at UNJ is expected to produce downstream benefits for pedagogical readiness.

5.7. Limitations

Several limitations should be acknowledged. First, the cross-sectional design precludes causal inference about the relationship between AI usage and geometry learning outcomes. Second, the analytic sample is substantially imbalanced, with UNM contributing 50.0% of respondents and UNJ 50.0% ($n = 89$). The balanced design ($n = 89$ per institution) was achieved through random subsampling of the larger UNM sample; while this ensures equal statistical power across institutions, it also reduces the effective UNM sample from 203 to 89, potentially discarding informative variance. Future studies should employ prospective stratified sampling to collect sufficiently large and balanced samples from each institution without the need for post-hoc subsampling, targeting $n \geq 120$ per institution to improve estimation precision and to provide adequate statistical power for structural equation modelling (SEM), which typically requires a minimum of 10 observations per free parameter (Faul et al., 2009), enabling future researchers to test the full five-factor AIGLS structure via confirmatory factor analysis. Third, the AIGLS in its current form contains exclusively positively worded items following the elimination of all negatively worded items, rendering it vulnerable to response set effects; future revisions should explore behavioural frequency formats or scenario-based items (Swain et al., 2008). Fourth, the study relies entirely on self-reported Likert ratings without behavioural verification through usage logs, introducing potential social desirability bias (Podsakoff et al., 2003). Fifth, the inclusion of Semester 2 students introduces some eligibility heterogeneity; future studies should implement more stringent course completion criteria and employ longitudinal designs to track AI usage trajectories across the full pre-service preparation programme.

6. Conclusions

This study introduced and validated the AI Usage in Geometry Learning Scale (AIGLS), a domain-specific psychometric instrument, and generated the first institution-comparative empirical portrait of AI usage in geometry learning among pre-service mathematics teacher students in Indonesia ($N = 178$; UNM $n = 89$, UNJ $n = 89$; overall $\alpha = 0.873$). Five principal conclusions emerge.

First, AI usage in geometry learning at both institutions has reached a level of normalisation that renders adoption-focused interventions obsolete. All five AIGLS dimensions fell within the High interpretive category ($M = 3.77\text{--}4.17/5.00$), and ChatGPT was near-

universally adopted (96.6%). Critically, a statistically significant institutional disparity in usage frequency was observed ($\chi^2(4) = 27.917, p < 0.001$), with 22.5% of UNM students reporting near-daily AI engagement compared to only 2.2% at UNJ. This pattern indicates that differences in habitual engagement are as consequential as differences in the quality of AI use. The policy priority must therefore shift from encouraging adoption to cultivating the criticality, purposefulness, and pedagogical intentionality with which AI is engaged.

Second, the five AIGLS dimensions are empirically distinguishable, with Critical Evaluation of AI (CEA) constituting a metacognitive competency structurally dissociated from the four behavioural usage dimensions ($\rho = 0.235\text{--}0.409$ vs. $\rho = 0.324\text{--}0.517$ among CUS, PSA, GV, and PR). Students may engage with AI intensively while remaining unable to detect errors in AI-generated outputs. This finding confirms that AI literacy does not emerge organically from use and requires explicit, structured scaffolding within geometry teacher preparation programmes.

Third, Geometric Visualisation is the dimension most sensitive to institutional context, exhibiting the largest institutional gap among all five dimensions ($r_{rb} = -0.38, p < 0.001$, medium effect). UNM students substantially outperformed their UNJ counterparts in AI-assisted geometric visualisation, a pattern that likely reflects differential access to visualisation-specific AI tools, differential pedagogical modelling by geometry faculty, and potentially differential access to premium AI tiers offering multimodal spatial reasoning capabilities.

Fourth, four of the five AIGLS dimensions—Problem-Solving Assistance, Pedagogical Readiness, Critical Evaluation of AI, and Conceptual Understanding Support (borderline, $p = 0.051$, non-significant after Bonferroni correction)—converge across institutions. This convergence suggests that the underlying determinants of these usage patterns operate through individual-level mechanisms such as prior AI exposure, general digital literacy, and self-regulation rather than through institutional affiliation. Accordingly, these dimensions warrant uniform national-level curricular responses applicable across all LPTKs.

Fifth, the systematic underperformance of all five negatively worded items (corrected item-total correlations $r_{it} = -0.07$ to 0.11) makes a contribution to the psychometric literature on AI usage measurement. Social desirability and acquiescence reversal effects appear particularly acute in instruments targeting academically sensitive AI behaviours, informing future instrument developers toward behavioural frequency or scenario-based item formats as more robust alternatives.

Taken together, the validated AIGLS provides a replicable instrument for future comparative, longitudinal, and intervention-based research on AI integration in geometry teacher preparation. Priority directions include: prospective multi-institution studies with stratified sampling ($n \geq 120$ per site) adequate for confirmatory factor analysis; experimental designs relating AIGLS dimension profiles to measured geometry learning outcomes; and institution-targeted interventions addressing GV engagement at UNJ and AI critical literacy development at scale across the LPTK network.

Several limitations temper these conclusions. The cross-sectional design precludes causal inference, and the balanced analytic subsample ($n = 89$ per institution, seed = 42) may underrepresent the full variance of the larger UNM pool. Reliance on self-reported Likert ratings without behavioural verification through usage logs remains susceptible to social desirability bias. Future studies adopting mixed-methods designs, pairing AIGLS data with log-based usage evidence and qualitative classroom observation, would substantially strengthen the evidential base and deepen understanding of how AI usage patterns translate into geometry learning outcomes for pre-service mathematics teachers.

Author Contributions: Conceptualization, M.A.N., A.S. (Anny Sovia) and H.U.; methodology, M.A.N.; software, M.A.N.; validation, M.A.N., N.T., A.S. (Andi Syukriani) and A.S. (Anny Sovia);

formal analysis, M.A.N.; investigation, M.A.N., N.T. and A.S. (Andi Syukriani); data curation, M.A.N.; writing—original draft preparation, M.A.N.; writing—review and editing, N.T., A.S. (Andi Syukriani), A.S. (Anny Sovia) and H.U.; visualization, M.A.N.; supervision, A.S. (Anny Sovia) and H.U.; project administration, H.U. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by Universitas Negeri Makassar under the PNBP Postgraduate Research Scheme, grant number 524/DST/UN36.12/TU/2026. The research was financed through DIPA Universitas Negeri Makassar No. SP DIPA-139.03.2.693386/2026 of 1 December 2025, in accordance with the Decree of the Rector of Universitas Negeri Makassar No. 324/DST/UN36/HK/2026 of 11 March 2026.

Institutional Review Board Statement: This study was conducted in accordance with the Declaration of Helsinki and approved by the Institutional Ethics Committee of the Institute for Research and Community Service (LP2M), Universitas Negeri Makassar (Protocol No. 109/DST/UN36.12/TU/2026).

Informed Consent Statement: Informed consent was obtained from all participants prior to their involvement in the study. Participants were informed that participation was entirely voluntary and that they could withdraw at any time without consequence.

Data Availability Statement: The data that support the findings of this study are available from the corresponding author upon reasonable request. The data are not publicly available as they contain information that could compromise the privacy of research participants.

Acknowledgments: The author gratefully acknowledges the cooperation of the Mathematics Education faculty coordinators at Universitas Negeri Makassar and Universitas Negeri Jakarta, and the pre-service mathematics teacher students who voluntarily participated in this study.

Conflicts of Interest: The authors declare no conflicts of interest.

References

- Abubakar, S., Jeilani, A., & Yusuf, M. (2025). The role of over-reliance on AI in the negative consequences of student learning: The moderating effects of ethical concerns and institutional policies. *Cogent Education*, 12(1), 2591503. [\[CrossRef\]](#)
- Anderson, L. W., & Krathwohl, D. R. (2001). *A taxonomy for learning, teaching, and assessing: A revision of bloom's taxonomy of educational objectives*. Longman.
- Awai, A., Naufal, M. A., Sutamrin, S., & Huda, M. (2024). Enhancing geometry achievement in pre-service mathematics teachers: The impact of a scaffolded flipped classroom using a learning management system. *Journal of Ecohumanism*, 3(6), 637–645. [\[CrossRef\]](#)
- Awai, A., Sutamrin, S., Ihsan, H., Nasir, N., Naufal, M. A., & Ikram, M. (2025). Spatial geometry learning in higher education through flipped classroom with metacognitive scaffolding: A mixed-methods study. *Online Learning in Educational Research (OLER)*, 5(2), 583–600. [\[CrossRef\]](#)
- Barnette, J. J. (2000). Effects of stem and Likert response option reversals on survey internal consistency: If you feel the need, there is a better alternative to using those negatively worded stems. *Educational and Psychological Measurement*, 60(3), 361–370. [\[CrossRef\]](#)
- Bastani, H., Bastani, O., Sungu, A., Ge, H., Kabakçı, Ö., & Mariman, R. (2025). Generative AI without guardrails can harm learning: Evidence from high school mathematics. *Proceedings of the National Academy of Sciences*, 122(26), e2422633122. [\[CrossRef\]](#) [\[PubMed\]](#)
- Battista, M. T. (1990). Spatial visualization and gender differences in high school geometry. *Journal for Research in Mathematics Education*, 21(1), 47–60. [\[CrossRef\]](#)
- Battista, M. T. (2007). The development of geometric and spatial thinking. In F. K. Lester (Ed.), *Second handbook of research on mathematics teaching and learning* (pp. 843–908). Information Age Publishing.
- Bryman, A. (2016). *Social research methods* (5th ed.). Oxford University Press.
- Celik, I., Dindar, M., Muukkonen, H., & Järvelä, S. (2022). The promises and challenges of artificial intelligence for teachers: A systematic review of research. *TechTrends*, 66(4), 616–630. [\[CrossRef\]](#)
- Chiu, T. K. F. (2025). Developing intelligent-TPACK (I-TPACK) framework from unpacking AI literacy and competency: Implementation strategies and future research direction. *Interactive Learning Environments*, 33(7), 4189–4192. [\[CrossRef\]](#)
- Clements, D. H., & Battista, M. T. (2006). Geometry and spatial reasoning. In D. A. Grouws (Ed.), *Handbook of research on mathematics teaching and learning: A project of the national council of teachers of mathematics* (pp. 420–464). Emerald Publishing Limited. [\[CrossRef\]](#)
- Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1), 37–46. [\[CrossRef\]](#)
- Cortina, J. M. (1993). What is coefficient alpha? An examination of theory and applications. *Journal of Applied Psychology*, 78(1), 98–104. [\[CrossRef\]](#)

- Creswell, J. W., & Creswell, J. D. (2023). *Research design: Qualitative, quantitative, and mixed methods approaches* (6th ed.). SAGE Publications, Inc.
- DeVellis, R. F., & Thorpe, C. T. (2021). *Scale development: Theory and applications*. SAGE Publications.
- Dilling, F., & Herrmann, M. (2024). Using large language models to support pre-service teachers mathematical reasoning—An exploratory study on ChatGPT as an instrument for creating mathematical proofs in geometry. *Frontiers in Artificial Intelligence*, 7, 1460337. [CrossRef] [PubMed]
- Drushlyak, M., Lukashova, T., Shamonia, V., & Semenikhina, O. (2025). ChatGPT-based simulation helps to develop the pre-service mathematics teachers' critical thinking. *International Journal of Instruction*, 18(1), 153–172. [CrossRef]
- Dunn, O. J. (1961). Multiple comparisons among means. *Journal of the American Statistical Association*, 56(293), 52–64. [CrossRef]
- Ergene, O., & Ergene, B. C. (2025). AI ChatBots' solutions to mathematical problems in interactive e-textbooks: Affordances and constraints from the eyes of students and teachers. *Education and Information Technologies*, 30(1), 509–545. [CrossRef]
- Etikan, I. (2016). Comparison of convenience sampling and purposive sampling. *American Journal of Theoretical and Applied Statistics*, 5(1), 1–4. [CrossRef]
- Facione, P. A. (1990). *Critical thinking: A statement of expert consensus for purposes of educational assessment and instruction*. California Academic Press. Available online: <https://eric.ed.gov/?id=ED315423> (accessed on 18 May 2026).
- Faul, F., Erdfelder, E., Buchner, A., & Lang, A.-G. (2009). Statistical power analyses using G*Power 3.1: Tests for correlation and regression analyses. *Behavior Research Methods*, 41(4), 1149–1160. [CrossRef] [PubMed]
- Hau, N. H., Nam, P. S., Son, T. C., Anh, D. C. L., Van, N. T., Tu, P. T. T., Nga, T. T., & Mai, V. X. (2026). Integrating generative ai and cultural storytelling to enhance geometry learning in Vietnamese primary classrooms: A quasi-experimental study. *Education Sciences*, 16(4), 588. [CrossRef]
- Kasneci, E., Sessler, K., Küchemann, S., Bannert, M., Dementieva, D., Fischer, F., Gasser, U., Groh, G., Günemann, S., Hüllermeier, E., Krusche, S., Kutyniok, G., Michaeli, T., Nerdel, C., Pfeffer, J., Poquet, O., Sailer, M., Schmidt, A., Seidel, T., . . . Kasneci, G. (2023). ChatGPT for good? On opportunities and challenges of large language models for education. *Learning and Individual Differences*, 103, 102274. [CrossRef]
- Lawshe, C. H. (1975). A quantitative approach to content validity. *Personnel Psychology*, 28(4), 563–575. [CrossRef]
- Lei, S., & Smith, M. R. (2003). Evaluation of several nonparametric bootstrap methods to estimate confidence intervals for software metrics. *IEEE Transactions on Software Engineering*, 29(11), 996–1004. [CrossRef]
- Li, D., Osman, S., Alhassora, N. S. A., Kumar, J. A., & Husain, S. K. S. (2025). Empowering K-12 mathematics teachers with artificial intelligence: A systematic review of insights, applications, and challenges. *IEEE Access*, 13, 117198–117209. [CrossRef]
- Liu, B., Zhang, W., & Wang, F. (2026). Can generative artificial intelligence effectively enhance students' mathematics learning outcomes?—A meta-analysis of empirical studies from 2023 to 2025. *Education Sciences*, 16(1), 140. [CrossRef]
- Long, D., & Magerko, B. (2020). What is AI literacy? Competencies and design considerations. In *Proceedings of the 2020 CHI conference on human factors in computing systems* (pp. 1–16). Association for Computing Machinery. [CrossRef]
- Ma, W., Adesope, O. O., Nesbit, J. C., & Liu, Q. (2014). Intelligent tutoring systems and learning outcomes: A meta-analysis. *Journal of Educational Psychology*, 106(4), 901–918. [CrossRef]
- Mbatha, M., & Bansilal, S. (2023). Using the van Hiele theory to explain pre-service teachers' understanding of similarity in Euclidean geometry. *Education Sciences*, 13(9), 861. [CrossRef]
- Mishra, P., & Koehler, M. J. (2006). Technological pedagogical content knowledge: A framework for teacher knowledge. *Teachers College Record: The Voice of Scholarship in Education*, 108(6), 1017–1054. [CrossRef]
- Naufal, M. A., Abdullah, A. H., Osman, S., Abu, M. S., & Ihsan, H. (2021). The effectiveness of infusion of metacognition in van Hiele model on secondary school students' geometry thinking level. *International Journal of Instruction*, 14(3), 535–546. [CrossRef]
- Naufal, M. A., Atan, N. A., Abdullah, A. H., & Abu, M. S. (2017). Problem solving, based on metacognitive learning activities, to improve Mathematical reasoning skills of students. *Man in India*, 97(12), 213–220.
- Naufal, M. A., Ihsan, H., Samsuddin, A. F., Zainal, Z., Ashari, N. W., & Hassan, M. N. (2025). Evaluating the geometric thinking levels of generation Z pre-service mathematics teachers in Indonesia. *International Journal of Evaluation and Research in Education (IJERE)*, 14(2), 918–925. [CrossRef]
- Nguyen, D. T., & Pham, Q. V. (2025). The evolving landscape of AI integration in mathematics education: A systematic review of trends (2015–2025). *Eurasia Journal of Mathematics, Science and Technology Education*, 21(10), em2714. [CrossRef] [PubMed]
- Ning, Y., Zhang, C., Xu, B., Zhou, Y., & Wijaya, T. T. (2024). Teachers' AI-TPACK: Exploring the relationship between knowledge elements. *Sustainability*, 16(3), 978. [CrossRef]
- Nunnally, J. C. (1978). *Psychometric theory*. McGraw-Hill.
- Podsakoff, P. M., MacKenzie, S. B., Lee, J.-Y., & Podsakoff, N. P. (2003). Common method biases in behavioral research: A critical review of the literature and recommended remedies. *Journal of Applied Psychology*, 88(5), 879–903. [CrossRef] [PubMed]
- Polit, D. F., & Beck, C. T. (2006). The content validity index: Are you sure you know what's being reported? Critique and recommendations. *Research in Nursing & Health*, 29(5), 489–497. [CrossRef] [PubMed]

- Polya, G. (1957). *How to solve it. A new aspect of mathematical method* (2nd ed.). Princeton University Press.
- Shulman, L. (1987). Knowledge and teaching: Foundations of the new reform. *Harvard Educational Review*, 57(1), 1–23. [CrossRef]
- Spearman, C. (1904). The proof and measurement of association between two things. *The American Journal of Psychology*, 15(1), 72–101. [CrossRef] [PubMed]
- Sutrisno, S., Azis, A., Setyawan, M. B., Dewi, D. A., Hidayah, Y., Hakiki, M., Hamid, M. A., & Fadli, R. (2025). Assessing students' readiness for artificial intelligence-based project learning to strengthen local wisdom values in Indonesia. *Cogent Education*, 12(1), 2582948. [CrossRef]
- Swain, S. D., Weathers, D., & Niedrich, R. W. (2008). Assessing three sources of misresponse to reversed Likert items. *Journal of Marketing Research*, 45(1), 116–131. [CrossRef]
- Tapan Broutin, M. S. (2024). Exploring mathematics teacher candidates' instrumentation process of generative artificial intelligence for developing lesson plans. *Yuksekokretim Dergisi*, 14(1), 165–176. [CrossRef]
- Tatto, M., Schwill, J., Senk, S., Ingvarson, L., Rowley, G., Peck, R., Bankov, K., Rodriguez, M., & Reckase, M. (2012). *Policy, practice, and readiness to teach primary and secondary mathematics in 17 countries. Findings from the IEA Teacher Education and Development Study in Mathematics (TEDS-M)*. International Association for the Evaluation of Educational Achievement (IEA).
- Turmuzi, M., Azmi, S., & Kertiyani, N. M. I. (2026). ChatGPT in school mathematics education: A systematic review of opportunities, challenges, and pedagogical implications. *Teaching and Teacher Education*, 170, 105286. [CrossRef]
- Ulusoy, F., & Girit-Yildiz, D. (2024). The quality of prospective mathematics teachers' dynamic geometry tasks in terms of the coordination between mathematical depth levels and technological actions. *Journal of Computer Assisted Learning*, 40(5), 2261–2277. [CrossRef]
- van Hiele, P. M. (1986). *Structure and insight: A theory of mathematics education*. Academic Press. Available online: <https://books.google.co.id/books?id=0sHaAAAAMAAJ> (accessed on 23 April 2026).
- VanLehn, K. (2011). The relative effectiveness of human tutoring, intelligent tutoring systems, and other tutoring systems. *Educational Psychologist*, 46(4), 197–221. [CrossRef]
- Walkington, C. (2026). The implications of generative artificial intelligence for mathematics education. *School Science and Mathematics*, 126, 356–365. [CrossRef]
- Wardat, Y., Tashtoush, M. A., AlAli, R., & Jarrah, A. M. (2023). ChatGPT: A revolutionary tool for teaching and learning mathematics. *Eurasia Journal of Mathematics, Science and Technology Education*, 19(7), em2286. [CrossRef] [PubMed]
- Weigand, H.-G., Hollebrands, K., & Maschietto, M. (2025). Geometry education at secondary level—A systematic literature review. *ZDM—Mathematics Education*, 57(4), 829–843. [CrossRef]
- Wijaya, T., Yu, Q., Cao, Y., He, Y., & Leung, F. (2024). Latent profile analysis of AI literacy and trust in mathematics teachers and their relations with AI dependency and 21st-century skills. *Behavioral Sciences*, 14, 1008. [CrossRef] [PubMed]
- Wilson, F. R., Pan, W., & Schumsky, D. A. (2012). Recalculation of the critical values for Lawshe's content validity ratio. *Measurement and Evaluation in Counseling and Development*, 45(3), 197–210. [CrossRef]
- Xiao, C., Ye, J., Esteves, R. M., & Rong, C. (2016). Using Spearman's correlation coefficients for exploratory data analysis on big dataset. *Concurrency and Computation: Practice and Experience*, 28(14), 3866–3878. [CrossRef]
- Zawacki-Richter, O., Marín, V. I., Bond, M., & Gouverneur, F. (2019). Systematic review of research on artificial intelligence applications in higher education—Where are the educators? *International Journal of Educational Technology in Higher Education*, 16(1), 39. [CrossRef]
- Zimmerman, B. J. (2000). Self-efficacy: An essential motive to learn. *Contemporary Educational Psychology*, 25(1), 82–91. [CrossRef] [PubMed]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.