International Journal of Social Science and Human Research

ISSN (print): 2644-0679, ISSN (online): 2644-0695

Volume 08, Issue 12 December 2025

DOI: 10.47191/ijsshr/v8-i12-98, Impact factor- 8.007

Page No: 10018-10026

Evaluating Teacher AI Literacy Training and Mentoring Outcomes: A Nonparametric Pretest–Posttest Study with Effect Size and Normalized Gain

Kartika Sari[1], Luh Putu Ida Harini[2], Ni Ketut Tari Tastrawati[3], Ni Luh Putu Suciptawati[4], Ni Putu Dinda Ayu Kertiani[5], Ni Made Dwija Pratiwi[6]

[1,2,3,4,5,6]Mathematics Department, Udayana University, Badung, Bali, Indonesia

ABSTRACT: This study aims to examine the effectiveness of an artificial intelligence (AI) literacy training and mentoring program in enhancing teachers' AI-related competencies. A quantitative one-group pretest–posttest design was employed, involving 25 junior high school teachers who participated in a three-day training program focusing on fundamental AI concepts and their practical application in instructional planning, assessment, and learning content development. Data were collected using standardized pretest and posttest instruments and analyzed through descriptive statistics, nonparametric inferential tests, effect size estimation, and normalized gain analysis. The results revealed a statistically significant improvement in posttest scores compared to pretest scores, indicating meaningful learning gains following the training. The magnitude of improvement was substantial, as reflected by a very large effect size and a mean normalized gain of 0.672, suggesting a medium category of achievable learning improvement. Furthermore, no statistically significant differences in learning gains were found between male and female teachers, indicating that the training outcomes were broadly equitable across gender groups. This is confirmed by effect size and the normalized gain values obtained which are almost the same in male and female.

KEYWORDS: Artificial intelligence literacy; teacher professional development; effect size; normalized gain

I. INTRODUCTION

Artificial intelligence (AI) technology is currently developing rapidly and can be applied to various fields, including education. In education, the emergence of generative AI systems, such as ChatGPT and other large language models, has attracted attention due to their potential to help teachers develop teaching materials, provide feedback, and manage learning tasks more efficiently (Holmes et al., 2019; Kasneci et al., 2023; Yang et al., 2023). Consequently, through careful integration, these technologies contribute to creating more adaptive and personalized learning environments, while reducing the administrative burden on teachers (Luckin et al., 2016; Chen et al., 2020; Onesi-Ozigagun et al., 2024).

While AI has potential applications in education, its benefits depend heavily on teachers' AI literacy. This is because AI output is not inherently reliable, neutral, or pedagogically appropriate. Furthermore, the uncritical use of AI has the potential to create inaccuracies, bias, and ethical issues in the learning process (Floridi et al., 2018; Williamson & Eynon, 2020; Long & Magerko, 2020). Therefore, teachers need not only to have an operational understanding of AI tools but also to develop conceptual understanding, critical awareness, and ethical considerations regarding the application of AI in education. Without adequate AI literacy, AI integration risks weakening the quality of learning. This is what happened at a junior high school in Bebandem Regency, Karangasem, Bali. The school, which implements the Merdeka Curriculum, consists of 35 civil servant teachers and 6 honorary teachers. Based on the results of an initial situation analysis, only 8.33% of the teachers were highly knowledgeable and had already applied AI in teaching, 41.67% were moderately knowledgeable, and 50% had little to no understanding of AI concepts and their applications. Therefore, AI literacy training and mentoring were conducted for teachers. Despite the increasingly prominent role of AI in education, several studies indicate that many teachers remain unprepared for meaningful and responsible AI adoption. Most previous studies on the application of AI in education have focused on theoretical discussions, the technology's potential, or educators' perceptions of its use (Luckin et al., 2016; Floridi et al., 2018; Holmes et al., 2019; Chen et al., 2020; Williamson & Eynon, 2020; Long & Magerko, 2020; Kasneci et al., 2023; Yang et al., 2023; Onesi-Ozigagun et al., 2024). Meanwhile, empirical research quantitatively evaluating the impact of structured AI literacy training and mentoring programs on teachers remains limited, particularly in terms of objectively measuring learning improvement following professional development interventions.

Learning improvement can be measured through statistical tests. Indicators that reflect the magnitude or practical relevance of observed changes can also be used. Effect sizes, such as Cohen's d, provide important information about the strength of an intervention, while normalized gain provides an indication of the proportion of potential learning gains achieved by participants relative to their initial competencies (Cohen, 1980; Hake, 1998; Lakens, 2013). Most empirical studies of learning gains only use statistical tests. Meanwhile, the combined application of effect size estimation and normalized gain analysis is still rarely used in research related to measuring effectiveness.

Furthermore, the issue of gender-related equity in learning outcomes in professional development related to AI literacy has received relatively little attention. Although gender differences in digital competency and technology adoption have been reported in several previous studies, these findings are inconsistent and highly context-dependent (van Deursen & van Dijk, 2014; Scherer et al., 2019; OECD, 2021). Therefore, it remains unclear whether short-term AI literacy training and mentoring programs can produce comparable learning gains for male and female teachers. Therefore, empirical research that explicitly examines gender-based learning outcomes using appropriate statistical techniques is urgently needed.

In response to this gap and considering the real-world situation in secondary schools as mentioned above, this study examines the effectiveness of an AI literacy training and mentoring program for teachers using a quantitative pretest–posttest design. This study aims to (1) evaluate changes in teachers' AI literacy competencies after the training, (2) measure the magnitude of learning improvements through effect sizes and normalized gain, and (3) exploring potential differences in learning improvement based on gender groups

II. METHODS

2.1 Research Design

This study employed a quantitative one-group pretest–posttest design to obtain preliminary evidence regarding changes in teachers' AI literacy competencies following participation in an AI literacy training and mentoring program. All participants completed a pretest prior to the intervention and a posttest after the training was completed. This design was selected to allow for an initial evaluation of learning improvement associated with the training program. However, as no control or comparison group was included, the findings are interpreted as associational rather than causal.

2.2 Participants

The participants consisted of 25 junior high school teachers, including 12 males and 13 females, from a public school in Bali, Indonesia. All participants voluntarily took part in the three-day AI literacy training and mentoring program and completed both the pretest and posttest assessments. The relatively small and context-specific sample reflects the exploratory nature of the study and limits the generalizability of the findings beyond similar educational settings.

2.3 Training and Mentoring Program Description

The participants consisted of 25 junior high school teachers, including 12 males and 13 females, from a public school in Bali, Indonesia. All participants voluntarily took part in the three-day AI literacy training and mentoring program and completed both the pretest and posttest assessments. The relatively small and context-specific sample reflects the exploratory nature of the study and limits the generalizability of the findings beyond similar educational settings

2.4 Research Instrument

Data were collected using a standardized written test designed to assess participants' AI literacy, including knowledge of basic AI concepts and the ability to apply AI tools in instructional settings. Test scores ranged from 0 to 100, with higher scores indicating greater levels of AI literacy. The same instrument was administered as both the pretest and posttest to enable direct comparison of participants' performance before and after the training

2.5 Data Analysis Procedure

Data analysis was conducted in several stages. First, graphical representations were used to visualize individual changes in pretest and posttest scores. Second, descriptive statistics—including mean, standard deviation, median, minimum, and maximum values—were calculated to summarize participants' performance before and after the intervention.

Prior to inferential analysis, the Shapiro–Wilk test was applied to assess the normality of score distributions, given the sample size of fewer than 50 participants.

Because the normality assumption was not consistently satisfied, nonparametric statistical tests were employed. Changes between pretest and posttest scores were examined using the Wilcoxon signed-rank test, while differences in learning outcomes between male and female participants were analyzed using the Mann–Whitney U test.

a. Shapiro–Wilk Test [16](Hollander et al., 2013)

The Shapiro–Wilk test measures the correlation between the ordered sample values and the expected normal scores.

The test statistic is calculated using equation

$$W = \frac{(\sum_{i=1}^{N} a_i x_{(i)})^2}{\sum_{i=1}^{N}(x_{(i)} - \bar{x})^2} \qquad (1)$$

where: $x_{(1)} \le x_{(2)} \le \cdots \le x_{(n)}$ denote the ordered sample values (order statistics); $\bar{x}$ is the sample mean; and a_i are fixed coefficients that depend on the sample size n, derived from the expected values and covariance structure of the order statistics of a standard normal distribution; $0 < W \le 1$. Decision criteria: If $p \le \alpha$, reject H_0, indicating that the sample significantly deviates from a normal distribution and therefore cannot be assumed to follow a normal distribution. If the data are not normally distributed, then it continued with Wilcoxon signed–rank test. If the data are normally distributed, it continued with paired t test.

b. Wilcoxon signed–rank test (Wilcoxon, 1945)

The test statistic is calculated using equation $T = \min(T^+, T^-)$ (2)

where T^+ : sum of ranks for positive differences; T^-: sum of ranks for negative differences. Decision criteria: For a two-tailed test at significance level (α), reject H_0 if the computed p-value $\le \alpha$.

c. Mann–Whitney U test (Hollander et al.,2013)

The test statistic is calculated using equation

$$U = \min(U_1, U_2) \qquad (3)$$

where $U_1 = n_1 n_2 + \frac{n_1(n_1+1)}{2} - R_1; U_2 = n_1 n_2 + \frac{n_2(n_2+1)}{2} - R_2$; R_1 = sum of the ranks belonging to Group 1; R_2 = sum of the ranks belonging to Group 2; n_1 = the sample size of Group 1; n_2 = the sample size of Group 2

Decision criteria: If $U < U_{critical}$, reject H_0

To complement significance testing, the magnitude of learning improvement was quantified using Cohen's effect size (d) for paired samples. In addition, the normalized gain (Hake's g) was calculated to assess the proportion of potential learning improvement achieved relative to participants' initial scores. The combined use of inferential statistics, effect size estimation, and normalized gain analysis was intended to provide a more comprehensive evaluation of training outcomes beyond p-values alone.

a. Cohen's d size is calculated using equation (Cohen, 1980)

$$d = \frac{\bar{d}}{SD_{\bar{d}}} \qquad (4)$$

where $\bar{d} = \frac{1}{N}\sum_{i=1}^{N} diff_i ; SD_{\bar{d}} = \sqrt{\frac{\sum_{i=1}^{N}(diff_i - \bar{d})^2}{N-1}} ; diff_i = post\ test\ score\ i - pre\ test\ score\ i$

Table 1. Cohen's d Category (Cohen,1980; Sawilowsky,2009)

d value	Category
0.2	Small
0.5	Medium
0.8	Large
>1.2	Very Large/Huge

b. Normalized gain is calculated using equation (Hake's *g*) (Hake, 1998)

$$g = \frac{\overline{po} - \overline{pr}}{100 - \overline{pr}} \qquad (5)$$

where $\overline{po}$: mean of posttest score; $\overline{pr}$: mean of pretest score. Then it is determined Hake's g category base on Table 2.

Table 2 Hake's g Category (Hake, 1998)

Hake'g value	Category
$g < 0.3$	Low gain
$0.3 \le g < 0.7$	Medium gain
$g \ge 0.7$	High gain

III. RESULT AND DISCUSSION

3.1 Result

The training and mentoring program on artificial intelligence literacy training dan mentoring was attended by 25 teachers from one of yunior high school in Bebandem Karangasem Bali Indonesia, that consisting of 12 male and 13 female participants. The activity was conducted over a three-day period, on July 2025 from 14 to 16. The effectiveness of the pogram was evaluated by examining changes in participants' pretest and posttest scores. The individual improvement in learning outcomes is illustrated in Figure 1, while Figure 2 presents the score distributions disaggregated by gender.

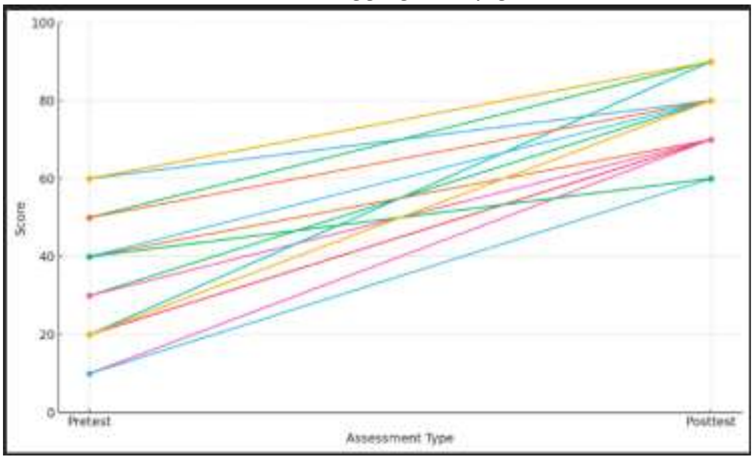

Figure 1 Individual Pretest -Posttest Improvement (Globally)

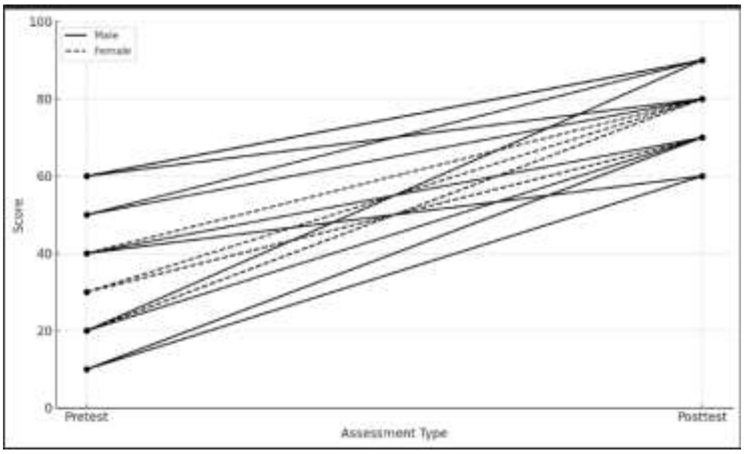

Figure 2 Individual Pretest -Postest Improvement By Gender

Figure 1 illustrates a consistent upward trend in all participants' scores from pretest to posttest, indicating that every individual demonstrated measurable improvement following the AI training program. The absence of any declining or static trajectories highlights the uniform effectiveness of the intervention across varying initial skill levels. Meanwhile, Figure 2 illustrates individual trajectories of pretest and posttest scores, differentiated by gender through solid (male) and dashed (female) line styles. Regardless of gender, all participants show clear upward score progressions, indicating consistent improvement following the training program. The uniform direction of increase, with no cases of score stagnation or decline, demonstrates the strong and comprehensive impact of the intervention across participants. This pattern suggests that this program enhanced participants' competencies in a gender-independent manner, reinforcing the overall robustness of the program's outcomes..

Although the figures illustrate general trends in participants' performance, descriptive statistical analysis is necessary to quantitatively characterize the distribution, central tendencies, and dispersion of the observed scores. Based on the 25 pretest and posttest scores, the results were summarized and presented in Table 3.

Table 3. Descriptive Statistic of Pretest and Postest Score of Responden

Statistic	N	Mean (M)	Standard Deviation (SD)	Minimum	25th Percentile	Median	75th Percentile	Maximum
Pretest Score	25	37.2	19.48	10	20	40	60	60
Posttest Score	25	79.2	10.38	60	70	80	90	90

Table 3 revealed a substantial increase in participants' performance from the pretest to the posttest following the training program. The pretest mean score (M = 37.20) fell within the low to moderate range, indicating that participants initially possessed limited skills related to the use of artificial intelligence in instructional practices. In contrast, the posttest mean score rose markedly to 79.20, demonstrating that the participants achieved a significantly higher level of competence after completing the training. Table 3 also show that the variability of scores also exhibited a notable shift. The standard deviation in the pretest (SD = 19.48) was considerably higher than that of the posttest (SD = 10.38), indicating that participants' abilities were more heterogeneous prior to the training but became more homogeneous afterward. This finding suggests that the training not only improved overall performance but also reduced disparities in participants' levels of mastery. Besides that, descriptive statistic analysis give that the distribution of scores further reinforces this pattern of improvement. The median pretest score of 40 increased to 80 in the posttest, indicating that more than half of the participants demonstrated at least a twofold enhancement in their initial proficiency. Additionally, the maximum posttest score reached 90, substantially higher than the pretest maximum of 60, suggesting that a subset of participants attained a remarkably high level of competence following the training.

In addition, descriptive statistics were generated separately for male and female participants to examine whether the observed improvements differed by gender, as presented in Table 4.

Table 4 Descriptive Statistic Analysis of Pretest dan Postest Score of Responden By Gender

Gender	Mean of Score		SD of Score		Min-Max Score		Median	
	Pretest	Posttest	Pretest	Posttest	Pretest	Posttest	Pretest	Posttest
Male (n=12)	37.50	78.33	21.79	11.93	10–60	60–90	45	80
Female (n=13)	36.92	80.00	17.97	9.13	10–60	60–90	30	80

Table 4 indicates that both male and femle participants experienced substantial improvement from pretest to posttest. The mean pretest scores for males (M = 37.50) and females (M = 36.92) were relatively comparable, suggesting that the two groups began the training with nearly equivalent initial proficiency levels. However, the distribution of male pretest scores exhibited greater variability (SD = 21.79) than that of females (SD = 17.97), indicating that the male participants' initial abilities were more heterogeneous. In the posttest results, both groups achieved the same median score of 80; however, female participants recorded a slightly higher mean score (M = 80.00) compared to their male counterparts (M = 78.33). Furthermore, the standard deviation among females (SD = 9.13) was lower than that of males (SD = 11.93), indicating that female participants demonstrated more consistent performance gains following the training. Base on these result, there was no evidence that either group derived systematically greater benefits from the intervention. These findings affirm that the artificial intelligence training program produced a broadly positive and inclusive impact, with no differential effects attributable to gender.

Although the graphical trends and descriptive statistics indicate an improvement in participants' competencies following the training, inferential statistical analysis was conducted to determine whether this improvement was statistically significant. As initial step in inferential statistical analysis is normality test.

A Given that the sample size was fewer than 50 participants, the Shapiro–Wilk test was employed to assess the distributional assumptions. Test statistic - W is calculated using (1) The results of the normality analysis are presented in Table 5.

Table 5 The result of Saphiro-Wilk Test

Variable	Statistic-W	p-value	Interpretasi
Pretest Score	0.8548	0.0022	Not normal distribution
Posttest Score	0.8455	0.0014	Not normal distribution

A normality assessment was performed on the pretest and posttest scors using the Shapiro–Wilk test, which is recommended for small to moderate sample sizes. Table 5 showed that both the pretest (W = 0.8548, p = 0.0022) and posttest scores (W = 0.8455, p = 0.0014) significantly deviated from a normal distribution. Because the p-values for both variables were below the significance threshold of 0.05, the null hypothesis of normality was rejected. These findings suggest that the distribution of participants' scores was not-normal, implying that subsequent statistical analyses should employ nonparametric methods that do not rely on the assumption of normality, yng dalam penelitian ini digunakan This distributional pattern is common in educational intervention studies, where initial abilities are often skewed and learning gains may cluster at higher values following training.

Because the normality assumption was not satisfied, the Wilcoxon signed–rank test was employed to examine differences between participants' pretest and posttest scores. Test statistic is calculated using (2). The analysis revealed a statistically significant increase in performance following the intervention ($W = 0.00$, $p < 0.001$). The extremely small p-value indicates that the probability of observing such differences by chance is negligble. All participants showed either improvement or no decrease, resulting in the lowest possible test statistic. These findings provide strong evidence that the training program substantially enhanced participants' competencies. The consistent upward shift in scores suggests that the intervention had a robust and uniform effect across respondents.

Subsequently, normality testing was carried out on the gender-based data to assess the distributional assumptions required for further statistical analysis. To examine whether the pretest and posttest scores met the assumptions required for parametric analysis, the Shapiro–Wilk test was conducted separately for male and female participants. Among male respondents, both the pretest ($W = 0.8126$, $p = 0.0131$) and posttest scores ($W = 0.8326$, $p = 0.0225$) significantly deviated from normality, indicating non-normal distributions. For female participants, the pretest scores did not significantly deviate from a normal distribution ($W = 0.8739$, $p = 0.0590$). However, the posttest scores were found to be non-normal ($W = 0.8557$, $p = 0.0339$). These results suggest that normality assumptions are not consistently satisfied across gender groups, particularly for post-intervention scores. Besause of these findings, nonparametric statistical procedures are more appropriate for subsequent analyses involving gender comparisons or within-group changes, as they do not rely on distributional assumptions. The presence of skewed distributions may reflect the strong ceiling effects observed after the training intervention, where many participants achieved similarly high posttest scores.

To determine whether learning outcomes differed between male and female participants, the Mann–Whitney U test was employed due to the non-normal distribution of several score variables. Test statistic U is calculated unsing (3) The analysis revealed no significant differences in pretest performance between males and females ($U = 76.5$, $p = 0.956$). The associated effect size was negligible ($r = 0.016$), indicating virtually no practical difference between the two groups prior to the intervention. Similarly, no significant gender differences were observed in posttest scores ($U = 74.0$, $p = 0.842$). The effect size for the posttest comparison was also negligible ($r = 0.044$), suggesting that both genders benefited from the intervention to a comparable extent. These findings imply that the training program was equally effective for male and female participants, without favoring any particular group.

Furthermore, the magnitude of change was evaluated using Cohen's effect size. To quantify the magnitude of the learning gains observed in the training program, Cohen's *d* was calculated using the paired-sample approach, which relies on the mean difference between pretest and posttest scores divided by the standard deviation of the difference scores. Cohen size is calculated using (4) and yielded a Cohen's *d* value of 2.91, which represents an exceptionally large effect according to conventional benchmarks. This magnitude indicates that the training intervention produced a substantial and highly meaningful improvement in participants' competencies. An effect size of this magnitude suggests that the posttraining performance was nearly three standard deviations higher than the pretraining performance, demonstrating the strong impact of the artificial intelligence training program on teachers' mastery of the targeted skills.

The effect size analysis using Cohen's d for paired samples indicated a very strong impact of the AI training program on both male and female participants. Male participants demonstrated a large improvement with an effect size of $d = 2.71$, while female participants exhibited an even stronger effect ($d = 3.00$). According to Cohen's classification(Table 1), these values represent extremely large effects, suggesting that the three-day training and mentoring program substantially enhanced teachers' competencies in AI-assisted lesson planning, presentation design, and instructional video production. The slightly higher effect size among female participants implies that they may have benefited more from the intervention, which warrants further exploration in future studies.

In addition to Cohen's effect size, the magnitude of improvement can also be assessed using normalized gain. To evaluate the overall instructional impact of the training program, the normalized gain (Hake's g) was calculated for each participant using (5). This metric measures the proportion of possible improvement achieved relative to the maximum attainable score. The analysis yielded an average normalized gain of $g = 0.672$, indicating a substantial improvement in participants' performance following the training. According to Hake's classification, normalized gains between 0.30 and 0.70 reflect medium learning effectiveness, while gains above 0.70 indicate high effectiveness. The observed value of 0.672 situates the training outcome at the upper end of the medium-effectiveness range, approaching the threshold for high effectiveness. This result suggests that the training program was highly successful in facilitating meaningful learning gains across participants. The consistently positive gains observed among all respondents also demonstrate the robustness of the instructional design and its broad applicability across participant groups.

The male participants achieved a normalized gain of $g = 0.653$, while the female participants obtained a slightly higher gain of $g = 0.683$. According to Hake's classification (Table 2), both values fall within the upper range of the 'medium-gain' category, indicating that the intervention produced substantial improvement in participants' competencies. The results suggest that the training was effective for both genders, with female participants showing a marginally stronger learning gain. These findings highlight the program's capacity to enhance teachers' skills in AI-supported lesson planning, presentation design, and instructional video development.

3.2 Discussion

The results of this study demonstrate an increase in teachers' AI literacy competencies after participating in the training and mentoring program. The statistically significant increase in post-test scores, accompanied by a very large effect size, provides strong empirical evidence that the intervention was effective in improving teachers' knowledge and practical understanding of AI-related concepts. The magnitude and consistency of the observed improvements (Figure 1) indicate that the training was strongly associated with increased meaningful learning.

From the perspective of adult learning theory, these results align with the principles of andragogy proposed by Knowles (1980), which emphasizes that adult learners benefit most from learning experiences that are problem-centered, relevant to their professional roles, and immediately applicable. The AI literacy training emphasized hands-on practice, contextual examples, and the direct application of AI tools to teacher tasks, such as lesson planning and instructional content creation. This practical orientation facilitated deeper cognitive engagement and supported effective knowledge transfer to participants' professional contexts.

Furthermore, the effectiveness of the intervention can be explained through the lens of social constructivism theory. According to Vygotsky's framework, learning occurs through social interaction and guided participation within the learner's zone of proximal development (Vygotsky, 1978). The mentoring activities in this program enabled participants to gradually transition from initial unfamiliarity with AI technology to independent and confident use. Collaborative discussions and guided exploration of the use of AI technology contributed to meaningful learning, strengthening conceptual understanding and skill acquisition.

The results also showed a significant reduction in individual pretest-to-posttest score variability, indicating increased homogeneity in participant competency after the training. This pattern suggests that the intervention not only improved overall performance but also helped reduce gaps among participants with varying initial skill levels. These findings align with previous research on differentiated professional learning, which emphasizes the importance of adaptive support and mentoring in addressing diverse learning needs among educators. By enabling participants to engage with AI tools at their own pace and receive targeted guidance, this training program reduced initial skill gaps and promoted more uniform learning outcomes. Such an approach reflects contemporary models of teacher professional development that prioritize inclusivity and personalized learning. (Castanheira, 2016; Allen et al., 2017; Boelens, et al., 2018; Aderibigbe et al., 2018; Vantieghem et al., 2020)

Furthermore, from a gender perspective, the study results showed no statistically significant differences in learning gains between male and female participants. This indicates that the AI literacy training and mentoring program was effective across gender lines. These findings support the perspective that AI literacy is not inherently gender-dependent but is shaped by exposure, instructional design, and learning support. The collaborative, mentoring-based training structure created an inclusive learning environment that enabled both male and female teachers to proactively engage in the use of AI technology (Younis, 2024; Samngamjan et al., 2024; MacDowell, 2024; Liao et al., 2025). These findings strengthen the argument that well-designed professional development programs can help address the perceived gender gap in emerging digital competencies.

The use of effect size estimates and normalized improvement analysis provides a more meaningful understanding of the impact of the training and mentoring intervention. The very large effect sizes obtained in this study indicate that the difference between pretest and posttest performance is substantial in practice, while the normalized improvement scores reflect the proportion of potential learning gains achieved by participants. These two measures indicate that the training was not only statistically effective but also educationally meaningful.

Ultimately, the results of this study have important implications for the development of AI literacy among educators. AI literacy extends beyond technical proficiency to include conceptual understanding, critical evaluation, and ethical awareness of AI systems. The observed improvements suggest that structured training programs combining coaching, hands-on practice, and mentoring can effectively support teachers in progressing from basic competency to functional competency in the use of AI. Within the broader framework of digital competency and 21st-century skills, the results of this study demonstrate the importance of integrating AI literacy into teacher professional development initiatives. Therefore, teachers must be equipped with the necessary knowledge and skills to ensure the responsible, effective, and pedagogically appropriate integration of AI into teaching and learning.

IV. CONCLUSIONS

This study provides empirical evidence suggesting that participation in an AI literacy training and mentoring program was associated with meaningful improvements in teachers' AI-related competencies. Participants demonstrated higher posttest scores compared to their pretest performance, supported by inferential analysis as well as effect size and normalized gain measures, indicating a substantial learning progression following the intervention. The results further suggest that learning gains were comparable across male and female teachers, indicating that the training supported equitable learning outcomes. While these findings point to the potential effectiveness of structured AI literacy training, they should be interpreted cautiously due to the absence of a control group and the limited sample size. Future studies employing more rigorous experimental designs and broader participant populations are recommended to strengthen the generalizability of these results.

ACKNOWLEDGMENT

This work was successfully carried out with the financial support of the *Program Udayana Mengabdi*, funded through the PNBP grant from Udayana University. The authors gratefully acknowledge this support, which made the completion of this activity possible.

REFERENCES

1) Aderibigbe, S., Gray, D., & Colucci-Gray, L. 2018. Understanding The Nature of Mentoring Experiences Between Teachers and Student Teachers. *International Journal of Mentoring and Coaching in Education, 7*(1), 54–71.
2) Allen, J.M., White, S., & Sim, C. 2017. Project Evidence: Responding to The Changing Professional Learning Needs of Mentors in Initial Teacher Education. *Australian Journal of Teacher Education, 42*(7), 14–25
3) Boelens, R., Voet, M., & De Wever, B. 2018. *The design of blended learning in response to student diversity in higher education: Instructors' views and use of differentiated instruction in blended learning. Computers & Education, 120*, 197–212. http://doi.org/10.1016/j.compedu.2018.02.009
4) Castanheira, P. S. P. 2016. Mentoring for Educators' Professional Learning and Development: A Meta-Synthesis of IJMCE Volumes 1–4. *International Journal of Mentoring and Coaching in Education, 5*(4), 334–346. https://doi.org/10.1108/IJMCE-10-2015-0030
5) Chen, L., Chen, P., & Lin, Z. 2020. Artificial intelligence in education: A review. *IEEE Access*, 8, 75264–75278. https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=9069875
6) Cohen, J. 1980. *Statistical Power Analysis for the Behavioral Sciences* (2nd ed.). United States of America: Lawrence Erlbaum Associates
7) Floridi L,Cowls J, Beltrametti M, Chatila R, Chazerand P, Dignum V, Luetge C, Madelin R, Pagallo U, Rossi F. Schafer B, Valcke P, Vayena E. 2018. AI4People—An Ethical Framework for A Good AI Society: Oportunities, Risk, Principles, and Recommendations. *Minds and Machines*, 28(4), 689–707.
8) Hake, R. R. 1998. Interactive-engagement Versus Traditional Methods. *American Journal of Physics*, 66(1), 64–74.
9) Hollander, M., Wolfe, D. A., & Chicken, E. 2013.*Nonparametric Statistical Methods* (3rd ed.). New York: John & Sons.
10) Holmes, W., Bialik, M., & Fadel, C. 2019. *Artificial Intelligence in Education: Promises and Implications for Teaching and Learning*. Center for Curriculum Redesign. https://curriculumredesign.org/wp-content/uploads/AIED-Book-Excerpt-CCR.pdf
11) Kasneci , E. Sessler , K., , Küchemann, S. , Bannert, M., Dementieva, D.,Fischer, F.,Gasser, U.,Groh, G., Günnemann, S., Hüllermeier, S. , Krusche, S. , Kutyniok, G. , Michaeli, T. , Nerdel, C., Pfeffer, J. ,Poquet , O. et al.. 2023. ChatGPT for Good? On Opportunities and Challenges of Large Language Models for Education. *Learning and Individual Differences*, 103, https://doi.org/10.1016/j.lindif.2023.102274
12) Knowles, M. S. (1980). *The Modern Practice of Adult Education: From Pedagogy to Andragogy*. Cambridge: The Adult Education Company.
13) Lakens, D. 2013. Calculating and Reporting Effect Size to Facilitate Cumulative Science: A Practical Primer for t-Test and ANOVAs. . *Frontiers in Psychology*, 4, 863. https://doi.org/10.3389/fpsyg.2013.00863
14) Liao, Z., Chen, Y & Mo, C. 2025. AI Literacy among Chinese Preschool Teachers: Empirical Status, Barriers and Cultivation Strategies. *Asian Journal of Education and Social Studies*, 51(7), 1351-1363. DOI 10.9734/ajess/2025/v51i72214
15) Long, D., & Magerko, B. (2020). What is AI literacy? Competencies and Design Considerations. *Proceedings of the CHI Conference on Human Factors in Computing Systems*. 1-16.
16) Luckin, R., Holmes, W., Griffiths, M., & Forcier, L. B. (2016). *Intelligence Unleashed: An Argument for AI in Education*. London: Pearson.
17) MacDowell, P., Moskalyk, K., Korchinski, K., & Morrison, D. 2024. Preparing Educators to Teach and Create With Generative Artificial Intelligence. *Canadian Journal of Learning and Technology.* 50(4) , 1-23.
18) OECD (2021), AI and the Future of Skills, Volume 1: Capabilities and Assessments, Educational Research and Innovation, OECD Publishing, Paris. https://doi.org/10.1787/5ee71f34-en.
19) Onesi-Ozigagun, O., Ololade, Y. J., Uyo-Udo, N.L., Ogundipe, D.O.2024. Revoluzionizing Education Through AI: A Comprehensive Review of Enhancing Learning Experiences. *International Journal of Applied Research in Social Sciences*, 6(4), 589-607.

20) Samngamjan, N., Phettom, P., Sa-ngunsat, K., & Philuek, W. 2024. A Survey Study on AI Literacy of Nakhon Sawan Rajabhat University's Digital Technology Teacher Students in Thailand. *Shanlax International Journal of Education, 13*(1), 33-40. https://doi.org/10.34293/education.v13i1.8355
21) Sawilowsky, S. S. 2009. New Effect Size Rules of Thumb. *Journal of Modern Applied Statistical Methods: 8*(2), 597–599.
22) Scherer, R., Siddiq, F., & Tondeur, J. 2019. The Technology Acceptance Model (TAM): A Meta-Analytic Structural Equation Modeling Approach to Explaining Teachers' Adoption of Digital Technology in Education. *Computers & Education*, 128, 13–35.
23) van Deursen, A. J., & van Dijk, J. A. (2014). The Digital Divide Shifts to Differences in Usage. *New Media & Society*, 16(3), 507–526.
24) Vantieghem, W., Roose, I., Gheyssens, E., Griful-Freixenet, J., Keppens, K., Vanderlinde, R., Struyven, K., & Van Avermaet, P. 2020. *Professional Vision of Inclusive Classrooms: A Validation of Teachers' Reasoning on Differentiated Instruction and Teacher-Student Interactions. Studies in Educational Evaluation, 67*, 100912. https://doi.org/10.1016/j.stueduc.2020.100912
25) Vygotsky, L. S. (1978). *Mind in Society: The Development of Higher Psychological Processes*. Cambridge: Harvard University Press
26) Wilcoxon, F. (1945). 1945. *Individual Comparisons by Ranking Methods.* Biometrics Bulletin. 1(6), 80–83. https://www.jstor.org/stable/i350470
27) Williamson, B., & Eynon, R. 2020. Historical threads, missing links, and future directions in AI in education. *Learning, Media and Technology*, 45(3), 223–235.
28) Yang, X., Wang, Q., & Lyu, J. 2023. *Assessing ChatGPT's Educational Capabilities and Application Potential. ECNU Review of Education*,7(3) , 699-713.
29) Younis, M. 2024. Effectiveness of a Professional Development Program Based on The Instructional Design Framework for AI Literacy in Developing AI literacy Skills Among Pre-Service Teachers. *Journal of Digital Learning in Teacher Education, 40*(2), 142–158. https://doi.org/10.1080/21532974.2024.2365663