JL3T (Journal of Linguistics, Literature and Language Teaching)
Vol. XI No. 2 (2025), pp. 213-231
https://doi.org/10.32505/jl3t.v11i2.12631

The Effectiveness of ChatGPT Integration in Improving English Learning Autonomy in Senior High Schools (SMA) in Banda Aceh City

Meta Keumala[1*], Zakaria[2]
English Language Education Department, Universitas Serambi Mekkah, Aceh, Indonesia
Economy Education Department, Universitas Serambi Mekkah, Aceh, Indonesia
✉ email: metakeumala@serambimekkah.ac.id

Received: 17 September 2025
Revised: 21 November 2025
Accepted: 01 December 2025

ABSTRACT

The rapid development of artificial intelligence (AI) has introduced new opportunities for language learning, particularly through tools such as ChatGPT. Despite its global use, the ChatGPT application in Indonesian high schools remains limited, particularly in supporting independent learning under the Independent Curriculum. This study investigated the effectiveness of ChatGPT in enhancing students' English learning independence, with a focus on engagement, motivation, and performance. A mixed-methods design was applied in three Banda Aceh schools (SMA 3, SMA 4, and SMA 5). Data were gathered through pre-tests and post-tests with 102 students divided into experimental groups using ChatGPT and control groups using conventional methods. Quantitative data were analysed using descriptive statistics, the Shapiro-Wilk test, Levene's test, and paired sample t-tests. Results showed significant improvement across schools. SMA 3 increased from a pretest mean of 80.39 to a posttest mean of 94.17 ($t = 16.60$, $p = 3.646e\text{-}18$). SMA 4 rose from 74.00 to 91.00 ($t = 23.49$, $p = 5.164e\text{-}23$), while SMA 5, with the lowest baseline (57.10), reached 83.93 ($t = 17.09$, $p = 1.106e\text{-}16$). Homogeneity tests indicated pretest variances differed ($p = 0.001$), but posttest variances became homogeneous ($p = 0.650$). The novelty of this study lies in showing that ChatGPT not only improved achievement but also reduced variability among students. In conclusion, ChatGPT effectively fostered independent English learning, balanced student outcomes, and offered a promising tool for adaptive, self-directed learning in Indonesian high schools.

Keywords: *Artificial intelligence; ChatGPT; ELT; Learning Autonomy; Senior high school*

Introduction

The rapid advancement of artificial intelligence (AI) technology has brought significant transformation across various fields, including education. In particular, recent developments in natural language processing (NLP) have enabled the emergence of highly sophisticated AI systems capable of understanding and generating human-like language Javaid, M., Haleem, A., Singh, R.P. and Suman, (2021). One of the most prominent innovations is ChatGPT, an NLP-based conversational agent developed to simulate interactive dialogue and provide context-sensitive responses. These capabilities make ChatGPT increasingly relevant in foreign language learning, where interaction, feedback, and exposure to authentic language use are essential. Research has shown that AI-powered language models can enhance learners' engagement, provide personalised

feedback, and facilitate more flexible learning experiences Dizon, G., & Tang, (2023); Dwivedi et al., (2023). In educational settings, ChatGPT has the potential to offer real-time explanations, answer students' questions, and support comprehension of complex learning materials, thereby functioning as an accessible and adaptive learning companion Kasneci et al., (2023). Through these affordances, ChatGPT contributes not only to improving language proficiency but also to promoting more autonomous and self-paced learning among students.

The integration of AI in education is expected to strengthen student independence by fostering learners' capacity to make autonomous decisions, regulate their learning behaviours, and sustain motivation throughout the learning process. In this context, independence refers to students' ability to manage their own learning cognitively and behaviourally—such as setting goals, selecting strategies, monitoring progress, and solving problems without excessive reliance on the teacher. Through adaptive feedback and personalised learning pathways, AI tools can promote these forms of autonomous learning, aligning with the 2025–2029 National Medium-Term Development Plan (RPJMN) of Indonesia, which underscores the importance of strengthening human resources through advancements in science, technology, and education Mulyasa, (2023).

Despite these opportunities, the application of AI, particularly ChatGPT, within the Indonesian educational context remains in its early stages of exploration Sari, A. R., & Setiawan, (2023). This limited adoption becomes especially significant in the context of the Independent Curriculum (Kurikulum Merdeka), which emphasises self-directed and personalised learning supported by technology Kemendikbudristek, (2022). However, the intended benefits of Kurikulum Merdeka are not yet fully realised in practice, as English learning continues to pose challenges for many high school students, including those in Banda Aceh Reza et al., (2023).

The role of AI in supporting personalised learning through adaptive feedback was emphasised; however, the focus was not on secondary school contexts Benvenuti et al., (2023). Meanwhile, the technical benefits of ChatGPT were discussed without examining its effects on students' self-regulation Sreen & Majid, (2024). The application of AI in self-regulated learning at the higher education level has been analysed, but the findings cannot be directly generalised to high school environments Chan et al., (2024); Chang et al., (2023); Mohebi, (2024).

While global research has widely examined the role of AI in education, studies specifically investigating the impact of ChatGPT on high school students' learning independence in Indonesia is underexplored. Several previous studies have discussed the benefits of AI-based learning tools. For instance, it was highlighted that chatbot-based learning can enhance student engagement in English classes; however, the research focused on simple chatbot applications without exploring their impact on learning independence AbuSahyon et al., (2023); Hakim, (2022); Perdana, (2024). Similarly, it was found that applications like Duolingo can increase learning motivation; nevertheless, the study did not address broader self-directed learning strategies such as goal-setting, monitoring progress, and reflecting on learning outcomes Muthohar et al., (2025); Qassrawi et al., (2024). The role of AI in supporting personalised learning

through adaptive feedback was emphasised; however, the focus was not on secondary school contexts, where students' autonomy is still developing and requires structured scaffolding Benvenuti et al., (2023).

Moreover, the technical benefits of ChatGPT—such as instant feedback and simplified explanations—were discussed without examining how these features influence students' self-regulation, metacognitive awareness, or independent problem-solving skills Sreen & Majid, (2024). Studies on the application of AI in self-regulated learning at the higher education level have also been conducted, but the findings cannot be directly generalised to high school environments because university students typically possess higher levels of learning autonomy and digital literacy Chan et al., (2024); Chang et al., (2023); Mohebi, (2024). These gaps show that the potential of ChatGPT to enhance learning autonomy—such as encouraging independent inquiry, reducing reliance on teacher explanation, and promoting reflective learning—remains underexplored, particularly within the context of Indonesian high schools implementing the Kurikulum Merdeka.

Building upon these gaps, this study contributes by examining how ChatGPT can enhance independent English learning among high school students in Banda Aceh. Specifically, this research seeks to investigate students' engagement, motivation, and development of independent learning strategies through AI-assisted interaction. The study also addresses local challenges, including limited technological infrastructure, varying levels of teacher and student readiness, and diverse cultural learning contexts.

Based on the research gaps outlined previously, the research questions are: (1) Does the use of ChatGPT impact high school students' English learning independence in Banda Aceh? and (2) Does the integration of ChatGPT increase students' motivation and engagement in learning English? Therefore, the objectives of this study are threefold: (1) to identify the effect of ChatGPT use on high school students' English learning independence in Banda Aceh and (2) to analyze the extent to which ChatGPT integration can increase student motivation and engagement.

Research Methods

The research methods section in this study provides a comprehensive account of the procedures undertaken, ensuring clarity and transparency for evaluation. The research method was descriptive quantitative conducted through quasi experimental design. According to Sukardi (2021), the quasi-experimental research method is considered one of the most productive approaches, because when conducted properly, it can address hypotheses that are primarily concerned with cause–effect relationships. The research began with a pre-test to establish baseline data on students' learning independence, followed by the implementation of ChatGPT in experimental groups and conventional methods in the control groups. Post-tests were then conducted to measure improvement, and the data were analysed using descriptive statistics, t-tests, and Pearson correlation coefficients in the SPSS software. This detailed and structured methodological design not only strengthens the validity of the findings but also allows readers to assess the appropriateness of the chosen methods for addressing the research objectives.

The population of this study consisted of all tenth-grade (Grade X) students. From this population, the sample was selected from regular classes, which were chosen because the students in these classes demonstrate relatively uniform academic abilities. This ensured that the sample reflected a balanced distribution of student competencies and allowed for more reliable comparison of learning outcomes within the experimental procedures. In addition, the instruments used in this study were pretest and posttest assessments covering two language skills: grammar and writing within the descriptive genre. The grammar test consisted of ten fill-in-the-blank items designed to measure students' comprehension of descriptive texts. The writing skill test required students to produce a simple three-paragraph descriptive composition. Each skill was evaluated using a distinct scoring rubric appropriate to the nature of the task, ensuring that both grammar and writing performance were assessed accurately and systematically.

Preparation

The activities in this stage include: (1) Literature study, (2) Preparation of Research Instruments, Instrument Validation, coordination with schools, ChatGPT usage training. To begin with, an in-depth literature review was conducted to support the theoretical basis of the research, particularly regarding the concept of independent learning in English language learning, the role of artificial intelligence in education, specifically ChatGPT, technology-based learning models under the Independent Curriculum, and Previous research related to AI and independent learning. This study strengthens the formulation of the problem and the design of the instrument.

Next, preparation of research instruments including pre-test and post-test instruments to measure changes in the level of student learning independence. Then, the instruments that have been prepared have been validated by experts in English language education and learning, as well as learning technology experts, to ensure the reliability and validity of the content. Moreover, the researcher coordinated with the schools, namely: SMA Negeri 3 Banda Aceh, SMA Negeri 4 Banda Aceh, and SMA Negeri 5 Banda Aceh. Last, the researchers conducted ChatGPT usage training to ensure that students could use ChatGPT optimally, researchers prepared a short training session (workshop) for the related experimental group.

Implementation

Pre-test

Implemented on two groups (class X-1 and X-2 in each school), namely the experimental group (using ChatGPT) and the control group (using conventional methods). The purpose of this action is to measure the initial level of student learning independence before treatment (baseline). Students in the experimental group used ChatGPT independently in English learning for 4 weeks (in the control class at each school). Activities include grammar practice, writing exercises, speaking prompts, and group discussions. The students were provided with structured, curriculum-based usage guidance.

Post-test

The post-test was conducted at the end of the experimental session (each school: 2 groups of grades 10) to measure the improvement of students' learning independence. The post-test results were compared with the pre-test using statistical tests.

Table 1 outlines the stages of data analysis conducted in the study, detailing the data types, analytical methods, and purposes of each step. The pre-test stage utilised descriptive statistics, such as mean, standard deviation, and range, to establish the baseline level of students' learning independence before the treatment.

Following the intervention, post-test data were analysed using both paired and independent t-tests to compare student performance before and after the treatment, as well as to assess the effectiveness of ChatGPT compared to conventional teaching methods. Overall, the analysis procedures presented in **Table 1** demonstrate a comprehensive mixed-methods approach, combining quantitative techniques to provide a holistic understanding of the impact of ChatGPT on student learning outcomes and experiences.

Table 1: Data Analysis Procedures

Stages	Data Types	Analysis Method	Purpose of Analysis
a. Pre-test	Learning independence score	Descriptive statistics (mean, SD, range) (Sukardi, 2013)	Measuring the initial level of student learning independence before treatment (baseline).
b. ChatGPT Implementation	ChatGPT usage activity	Qualitative descriptive (coding & thematic) (Takona, 2024)	Assess the intensity of engagement, the quality of student interaction with ChatGPT, and its suitability to the curriculum.
c. Post-test	Learning independence score	Paired t-test & Independent t-test (Field, 2024)	Testing the differences before and after treatment, as well as the effectiveness of ChatGPT compared to conventional methods.

Pre-Test

At the pre-test level, the type of data collected was students' learning independence scores. This data was analysed using descriptive statistics, including calculation of the mean, standard deviation, and range. The purpose of the analysis at this stage was to measure the initial level of students' learning independence before treatment, thus serving as a baseline for the research.

Action

Next, at the stage of ChatGPT implementation. The data obtained consisted of students' activities related to ChatGPT usage. The analysis was conducted descriptively and qualitatively using coding and thematic techniques. The aim was to assess the intensity of student engagement, the quality of interactions with ChatGPT, and the alignment of learning practices with the applicable curriculum.

Post-Test

At the post-test level, the reused data consisted of students' learning independence scores. Analysis was conducted using inferential statistical tests, namely paired t-tests and independent t-tests. The purpose of this stage was to examine differences in learning independence levels before and after treatment and to evaluate the effectiveness of using ChatGPT compared to conventional learning methods.

Results

a. Data Description

The descriptive statistics in **Table 2** provide an overview of the pretest and posttest performance of students from three schools (SMA 3, SMA 4, and SMA 5). At the pretest stage, SMA 3 recorded the highest average score, with a mean of 80.39, while SMA 5 had the lowest mean, at 57.10. The pretest standard deviation indicates that SMA 5 had the widest variation in student performance (SD = 12.00), suggesting a more heterogeneous group, whereas SMA 4's more minor deviation (SD = 6.68) shows a more homogeneous distribution of scores. Posttest results show an improvement across all schools, with SMA 3 maintaining the highest performance (Mean = 94.17). SMA 4 also showed a considerable increase, reaching a mean score of 91.00, while SMA 5, despite starting from a much lower baseline, improved significantly to 83.93. The increase in means across all schools indicates that the intervention or learning process conducted between pretest and posttest was effective in enhancing student outcomes.

Table 2: Pretest and Posttest Performance of Students from Three Schools

School	N	Min Pre	Max Pre	Mean Pre	SD Pre	Mean Post	SD Post
SMA 3	36	62	92	80.39	8.04	94.17	5.86
SMA 4	36	62	86	74.00	6.68	91.00	6.00
SMA 5	30	42	80	57.10	12.00	83.93	5.55

When examining variability, SMA 3 showed a decrease in standard deviation from 8.04 in the pretest to 5.86 in the posttest, suggesting that student performance became more consistent after the learning process. SMA 4 also displayed a slight reduction in variability (from 6.68 to 6.00), indicating stability in improvement. Similarly, SMA 5's deviation dropped from 12.00 to 5.55, reflecting that students not only improved on average but also became more uniform in their performance levels after the intervention. Overall, the data reveal positive learning gains across the three schools, with SMA 3 consistently leading in both pretest and posttest results. However, the most notable progress can be observed in SMA 5, which, despite its initially lower baseline, achieved a significant improvement in both mean score and reduced variability of student outcomes. These findings highlight the importance of considering both mean performance and standard deviation when evaluating the effectiveness of educational

interventions, as they provide insights into not only average improvement but also equity in learning outcomes across students.

b. Normality Test (Shapiro-Wilk)

The results of the Shapiro-Wilk normality test indicate that the pretest data for SMA 3 ($p = 0.092$) follow a normal distribution since the p-value is greater than 0.05. However, the posttest data for SMA 3 ($p = 0.000$) do not meet the assumption of normality. For SMA 4, both the pretest ($p = 0.035$) and posttest ($p = 0.008$) data fall below the 0.05 threshold, indicating that neither set of scores is usually distributed. Similarly, in SMA 5, the pretest ($p = 0.021$) and posttest ($p = 0.001$) data also fail the normality assumption. Overall, the findings suggest that most of the datasets, except the pretest scores of SMA 3, are not normally distributed. This implies that parametric statistical tests, which rely on the assumption of normality, may not be suitable for analysing the posttest scores across the three schools. Instead, non-parametric alternatives should be considered to ensure more accurate and reliable interpretations of the data. These results highlight the importance of verifying distributional assumptions before selecting appropriate statistical methods in educational research.

Normality tests were performed using the Shapiro-Wilk test. Data were considered normally distributed if $p > 0.05$.

- SMA 3: Pretest $p = 0.092$, Posttest $p = 0.000$
- SMA 4: Pretest $p = 0.035$, Posttest $p = 0.008$
- SMA 5: Pretest $p = 0.021$, Posttest $p = 0.001$

c. Homogeneity Test (Levene's Test)

The results of the homogeneity of variance test show that for the pretest scores, the p-value is 0.001, which is below the threshold of 0.05. This indicates that the variances of the pretest scores among the three schools are not homogeneous, meaning there are significant differences in the spread of scores before the intervention. Such a result suggests that the initial conditions of the students' performance levels varied widely across schools, which needs to be considered in interpreting subsequent learning outcomes. In contrast, the posttest results show a p-value of 0.650, which is greater than 0.05. This indicates that the variances of the posttest scores are homogeneous across the three schools. In other words, after the learning intervention, the distribution of student performance became more consistent between schools. This finding suggests that the intervention not only improved overall performance but also helped balance the differences in variability of scores among the schools.

Results of the homogeneity of variance test between schools:

- Pretest: $p = 0.001$
- Posttest: $p = 0.650$

d. Paired Sample T-Test

The results presented in **Table 3** indicate that the paired sample t-test revealed significant differences between pretest and posttest scores in all three schools. For SMA

3, the obtained t-value of 16.60 with a p-value of 3.646e-18 demonstrates a strong statistical significance, suggesting that the posttest scores were substantially higher than the pretest scores. In SMA 4, the difference was even more pronounced, with the highest t-value of 23.49 and an extremely small p-value of 5.164 × 10^ (-23), highlighting the most substantial learning gains among the three schools. Likewise, SMA 5 also showed significant improvement with a t-value of 17.09 and a p-value of 1.106e-16, confirming the effectiveness of the intervention in raising student performance.

Overall, **Table 3** clearly shows that the learning intervention had a significant positive impact on students' academic outcomes across the three schools. The consistently low p-values (all far below 0.05) validate the reliability of the results and confirm that the improvements were not due to chance. Notably, SMA 4 exhibited the most substantial impact of the intervention, while SMA 5 also achieved remarkable gains considering its lower baseline in the pretest. These findings reinforce the conclusion that the applied teaching strategy was successful in enhancing student achievement across different school contexts.

Table 3: Paired Sample T-Test Results

School	t-test	df	P=Value	Conclusion
SMA 3	16.60	35	3.646e-18	Significant
SMA 4	23.49	35	5.164e-23	Significant
SMA 5	17.09	29	1.106e-16	Significant

Discussion

The Q-Q plot presented in **Figure 1** illustrates the distribution of pretest scores from SMA 3 in relation to a theoretical normal distribution. Each blue dot represents an ordered value from the dataset, while the red line represents the expected normal distribution. When the data points fall closely along the red diagonal line, it suggests that the dataset approximates a normal distribution. In this figure, most of the data points align fairly well with the red line, particularly in the middle range of values, which indicates that the pretest scores are approximately normally distributed. Closer inspection reveals that while the majority of the points lie near the line, there are minor deviations at both the lower and upper ends of the distribution. These slight departures suggest a modest skewness or presence of outliers in the tails. However, such deviations are relatively small and do not drastically distort the overall pattern. This observation supports the results from the Shapiro-Wilk test, where the pretest data for SMA 3 showed a p-value greater than 0.05, confirming normality.

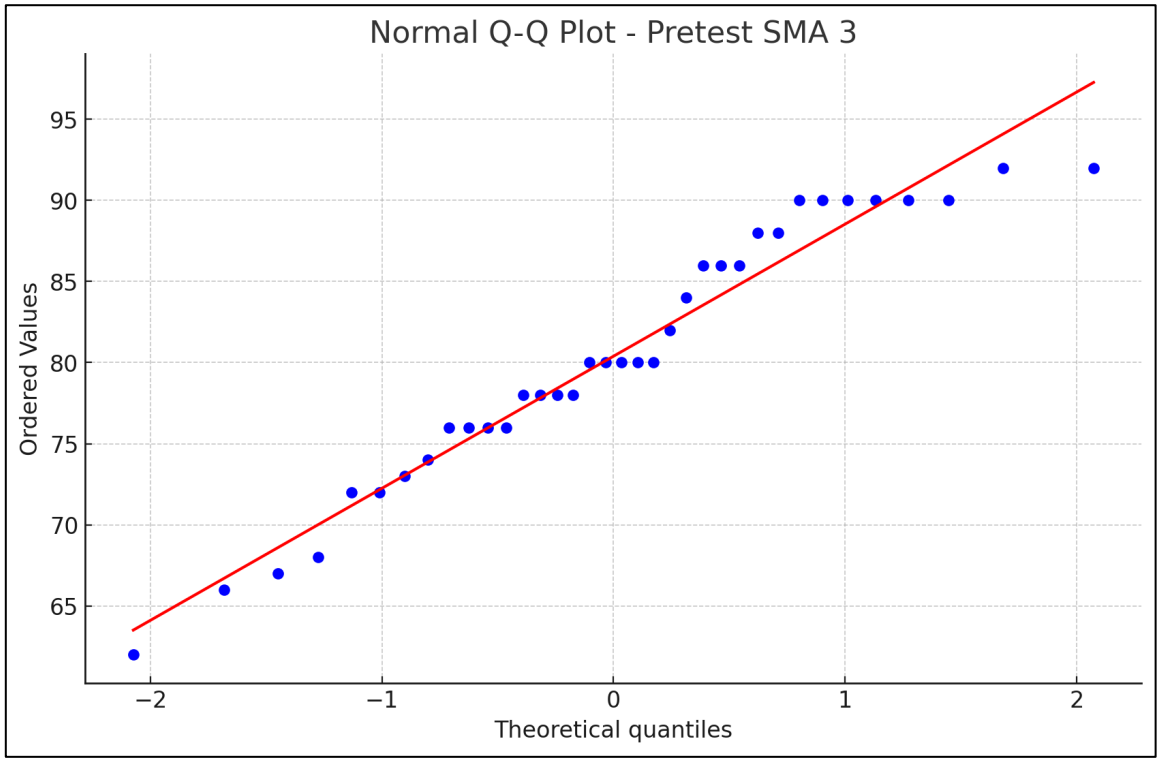

Figure 1: Normal Q-Q Plot of Pretest Scores in SMA 3

Another important aspect highlighted by the Q-Q plot is the consistency in the distribution of scores across the central range. The linear alignment of the majority of data points within this middle portion suggests that the majority of students' pretest scores followed a distribution pattern consistent with normality. This is particularly useful because the assumption of normality is often required when conducting parametric tests such as the t-test. The Q-Q plot thus provides strong visual confirmation that these statistical techniques are suitable for analysing this dataset. Overall, **Figure 1** offers valuable insight into the distribution of SMA 3's pretest data. While there are slight variations at the extremes, the general conformity of the points to the diagonal line suggests that the assumption of normality holds reasonably well. This makes the pretest data from SMA 3 suitable for further parametric statistical analysis, thereby strengthening the validity of the results obtained in subsequent tests, such as the paired sample t-test. The visual evidence from the Q-Q plot complements the statistical test results, offering a more comprehensive understanding of the data distribution.

The Q-Q plot in **Figure 2** illustrates the distribution of posttest scores from SMA 3 compared to a theoretical normal distribution. Each blue dot represents an observed data point, while the red line represents the expected values if the data followed a perfectly normal distribution. Ideally, if the data were normally distributed, the points would fall along the red diagonal line. However, in this figure, there is an apparent deviation, particularly at the upper end of the distribution where many data points cluster and flatten near the maximum score. A closer examination reveals that while the lower and middle portions of the distribution align somewhat with the red line, the upper section

diverges significantly. Many of the points at the higher end of the ordered values are above the line, reflecting a ceiling effect in the scores. This clustering suggests that a large number of students achieved very high posttest scores, which disrupts the symmetry of the distribution. As a result, the distribution deviates from normality, which is consistent with the Shapiro-Wilk test result for SMA 3 posttest (p = 0.000), indicating non-normality.

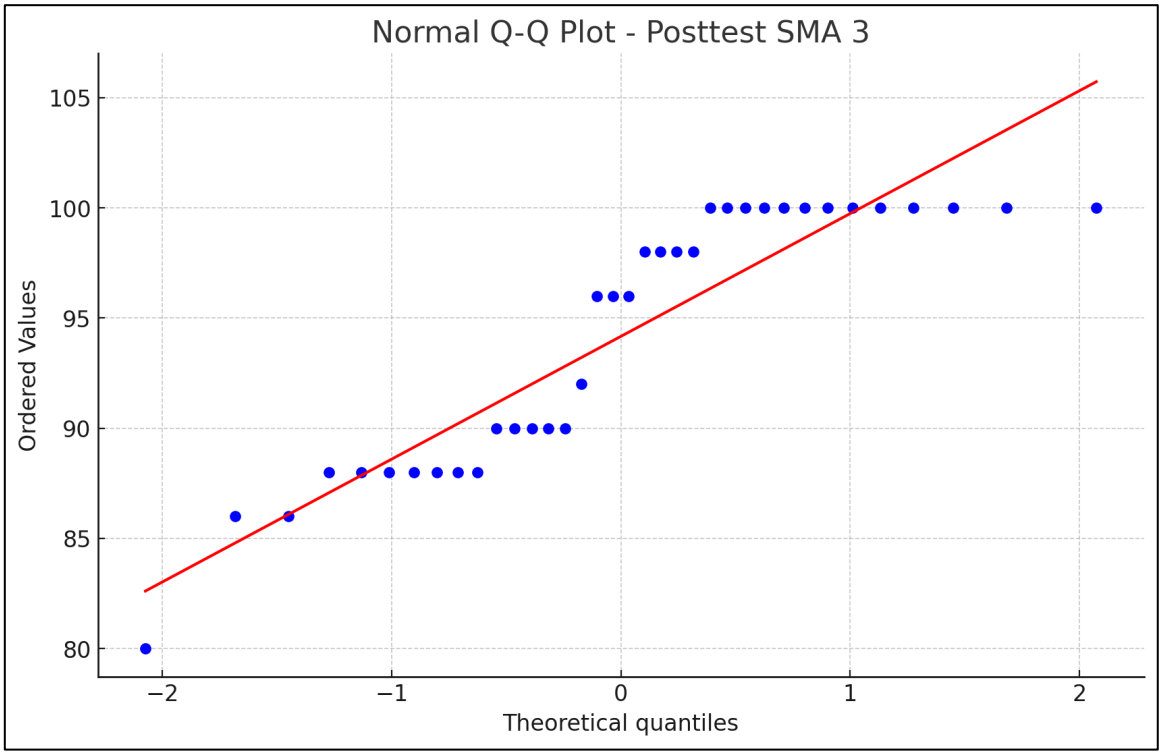

Figure 2: Normal Q-Q Plot of Posttest Scores in SMA 3

The pattern displayed in this Q-Q plot provides insight into the effectiveness of the intervention. The clustering of scores near the maximum suggests that students performed exceptionally well, with many reaching or approaching the highest possible scores. While this indicates success in terms of learning achievement, it also reduces variability among the scores, creating a skewed distribution. This is an essential consideration because it limits the applicability of parametric tests that assume normality, suggesting that non-parametric alternatives may be more appropriate for further analysis. In summary, **Figure 2** visually demonstrates that the posttest data for SMA 3 do not conform to a normal distribution. The ceiling effect, evident from the concentration of scores at the upper limit, not only explains the deviation from normality but also highlights the overall effectiveness of the teaching intervention in raising student performance. However, from a statistical perspective, these findings highlight the importance of carefully selecting analytical methods when evaluating posttest outcomes, ensuring that conclusions drawn remain valid despite any departure from normality.

The Q-Q plot shown in **Figure 3** presents the distribution of pretest scores from SMA 4 compared to a theoretical normal distribution. The blue dots represent the

ordered values of the observed data, while the red diagonal line represents the expected normal distribution. Ideally, if the data followed a perfectly normal distribution, the points would align closely with the red line. In this case, many of the points roughly follow the line in the central region, but deviations can be observed at both the lower and upper tails. In the lower tail of the distribution, several points fall below the red line, suggesting that there are lower scores than what would be expected in a normal distribution. Similarly, in the upper tail, the points tend to diverge slightly above the line, indicating a spread of higher scores. These deviations point to the possibility of skewness or non-normality in the distribution of SMA 4's pretest scores. This visual evidence is consistent with the Shapiro-Wilk test results, which showed a p-value of 0.035 for the pretest data of SMA 4, thereby rejecting the assumption of normality.

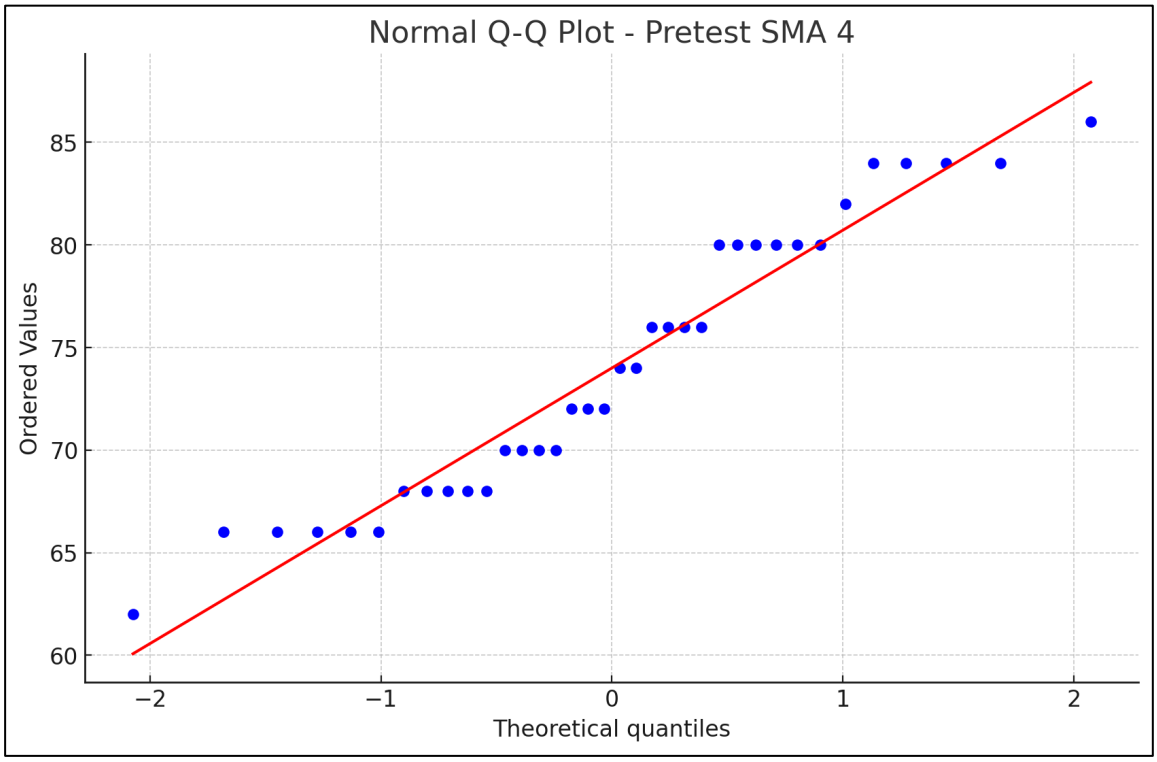

Figure 3: Normal Q-Q Plot of Pretest Scores in SMA 4

Despite these deviations, the central portion of the distribution aligns reasonably well with the theoretical line. This suggests that while the overall dataset may not perfectly follow a normal distribution, a substantial portion of the scores falls within an expected range. However, the departures in the tails are sufficient to disrupt normality, which is a critical consideration when selecting the appropriate statistical methods. Specifically, parametric tests that rely on the normality assumption may not be entirely relevant, and non-parametric alternatives might provide more robust results. Overall, **Figure 3** provides a clear visualisation of why SMA 4's pretest scores fail the normality test. The presence of deviations in both tails indicates a distribution that departs from normality, even though the middle section appears fairly aligned. This highlights the importance of combining visual tools, such as Q-Q plots, with statistical tests like the

Shapiro-Wilk test for a more comprehensive assessment of data distribution. The findings emphasise that researchers must carefully evaluate assumptions before conducting inferential analyses to ensure the validity of their conclusions.

The Q-Q plot shown in **Figure 4** illustrates the distribution of posttest scores for SMA 4 in comparison to a theoretical normal distribution. Each blue dot represents an observed data point, while the red line indicates the expected values under normality. At first glance, the data points appear to roughly follow the diagonal line in the middle range, suggesting some level of alignment with a normal distribution. However, deviations can be observed, particularly at the lower and upper ends, which indicate that the dataset does not fully adhere to normality. In the lower tail, several data points fall below the red line, reflecting that some of the lowest posttest scores are lower than expected under a normal distribution. Meanwhile, in the upper tail, a number of data points deviate above the line, showing a clustering of high scores near the maximum. This ceiling effect is a sign that many students achieved very high scores, which disrupts the balance of the distribution. These deviations confirm the findings from the Shapiro-Wilk test ($p = 0.008$), which also concluded that the posttest scores of SMA 4 are generally not distributed.

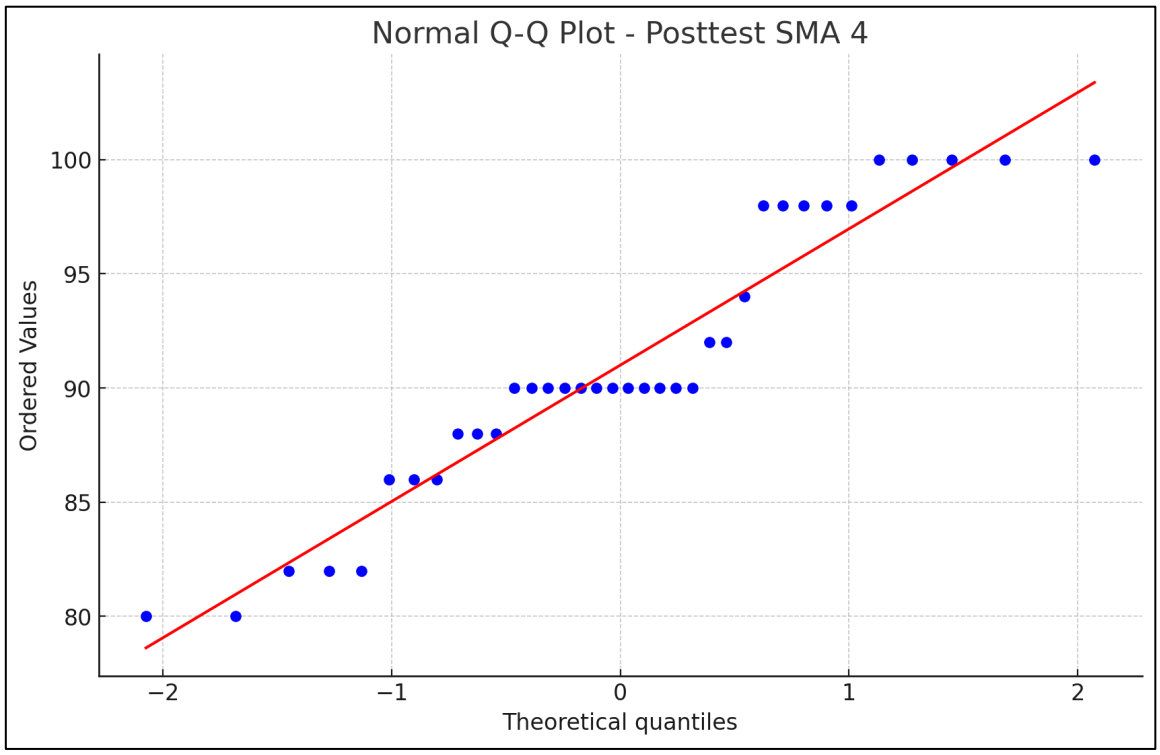

Figure 4: Normal Q-Q Plot of Posttest Scores in SMA 4

Despite the deviations, the alignment of points in the central portion of the plot indicates that the bulk of the scores fall relatively close to a normal distribution. This suggests that most students achieved scores within a consistent range, but the presence of both very high and somewhat lower scores introduced distortions at the tails. Such patterns are common in educational data when an intervention is highly effective,

resulting in many students clustering near the top end of the scale. Overall, **Figure 4** provides a clear visual representation of why SMA 4's posttest scores deviate from normality. While the middle range of data aligns reasonably well with the theoretical line, the tails deviate significantly due to clustering of high-performing students. This indicates substantial academic achievement following the intervention; however, it complicates statistical analysis, as the assumptions of normality are violated. As such, caution should be exercised when applying parametric tests, and non-parametric alternatives may be better suited for analysing the post-test performance of SMA 4 students.

The Q-Q plot in **Figure 5** illustrates the distribution of pretest scores for SMA 5 in comparison to a theoretical normal distribution. The blue dots represent the observed data values, while the red line indicates the expected values under normality. Ideally, if the data were perfectly normally distributed, the points would align closely along the diagonal line. At first glance, many of the data points appear to follow the line, particularly in the middle range; however, noticeable deviations are evident at the lower and upper tails. In the lower tail of the distribution, several points fall below the red line, suggesting that the lowest scores are lower than what would be expected under a normal distribution. Meanwhile, in the upper tail, the points slightly deviate above the line, showing that some of the higher scores are larger than expected. These deviations indicate that the dataset departs from a perfectly normal distribution, which is consistent with the Shapiro-Wilk test result for SMA 5 pretest ($p = 0.021$), confirming that the assumption of normality is violated.

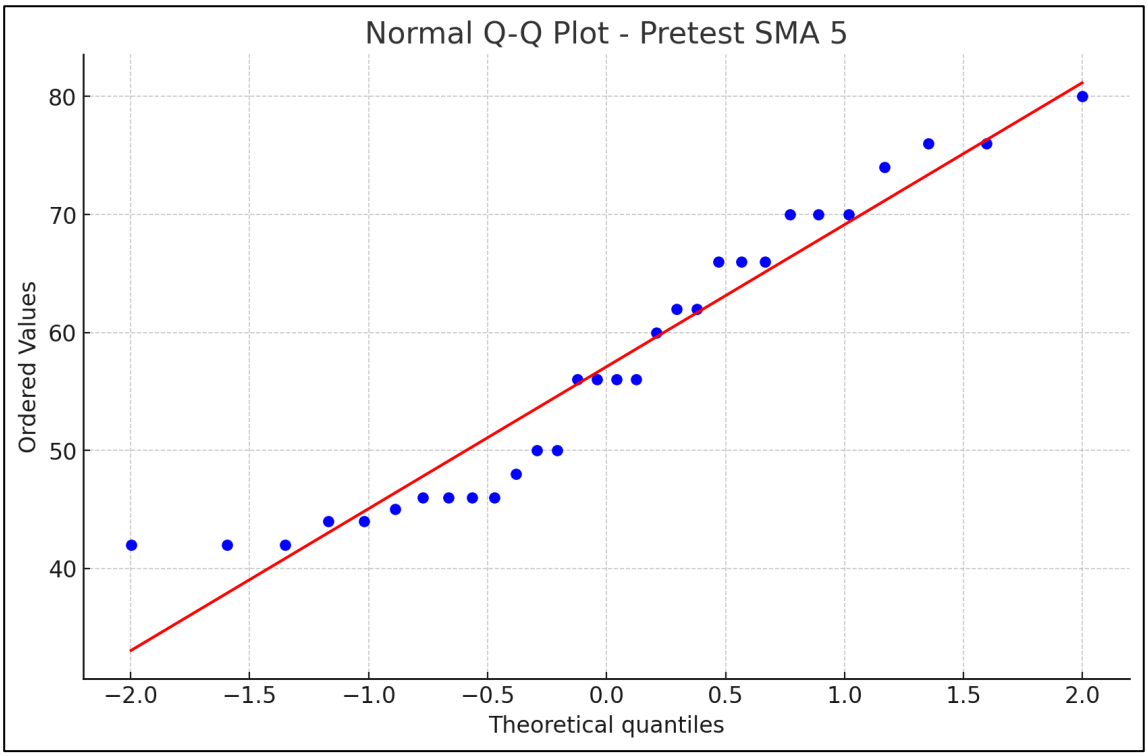

Figure 5: Normal Q-Q Plot of Pretest Scores in SMA 5

The central portion of the distribution, however, aligns relatively well with the theoretical line. This suggests that, while the extreme values diverge, a significant portion of the data is distributed close to a normal distribution. This pattern is typical in educational data where a diverse student population results in a wider spread of scores, particularly at the lower end. The presence of outliers or extreme values contributes to the departure from normality, making the distribution less symmetric and more irregular. Overall, **Figure 5** highlights that SMA 5's pretest scores do not strictly follow a normal distribution. The deviations at both ends of the distribution, especially at the lower tail, reflect variability in students' baseline performance before the intervention. These findings imply that parametric statistical tests may not be entirely appropriate for this dataset, and non-parametric alternatives should be considered. From an educational perspective, the plot also underscores the heterogeneity of student preparedness in SMA 5, with a broader spread of pretest scores compared to other schools.

The Q-Q plot in **Figure 6** displays the distribution of posttest scores from SMA 5 in comparison to a theoretical normal distribution. The blue points represent the observed scores, while the red diagonal line represents the expected values under a perfectly normal distribution. Ideally, the data points would align closely along the red line if the data followed a normal distribution. However, the figure reveals several deviations, particularly at both the lower and upper tails, indicating that the posttest scores for SMA 5 are generally not distributed. In the lower portion of the distribution, the points deviate slightly below the line, reflecting that some lower scores are lower than would be expected in a normal distribution. Notably, in the upper portion, the data points exceed the line, indicating that many students scored higher than expected. This clustering of points at the higher end reflects the effectiveness of the intervention, which enabled a majority of students to achieve strong posttest results. Such patterns, however, distort the normality of the distribution and are consistent with the Shapiro-Wilk test results ($p = 0.001$), which confirmed a violation of normality.

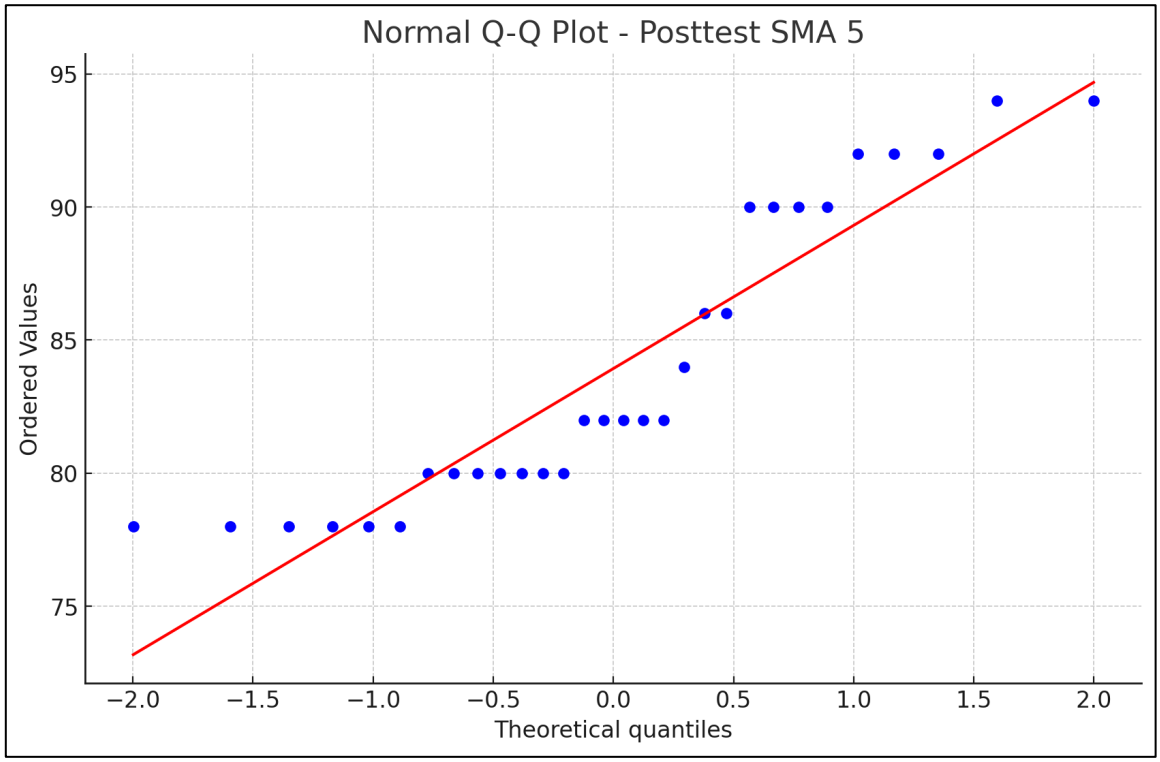

Figure 6: Normal Q-Q Plot of Posttest Scores in SMA 5

Although the tails deviate, the middle portion of the data aligns reasonably well with the diagonal line, suggesting that most scores are relatively consistent and fall within a predictable range. This indicates that while the data as a whole do not meet the assumption of normality, the majority of students' scores cluster within a central band. The improved consistency reflects the narrowing of score variation after the intervention, as indicated by the reduced standard deviation in descriptive statistics. Overall, **Figure 6** demonstrates that SMA 5's posttest distribution deviates from normality due to skewness and clustering at the higher end of the scale. This finding highlights both statistical and educational implications: statistically, it means that non-parametric tests may be more suitable for analysing these results, while educationally, it shows that the intervention was effective in significantly improving student performance. The ceiling effect observed at the top of the distribution underscores that many students reached high levels of achievement, reducing variability but also contributing to the departure from a standard curve.

The boxplot in **Figure 7** illustrates the distribution of posttest scores across SMA 3, SMA 4, and SMA 5, highlighting both improvements in mean performance and reductions in variability among students. SMA 3 achieved the highest overall scores, with most students clustered at the upper range, reflecting strong academic outcomes. Similarly, SMA 4 showed a substantial rise in performance, with scores concentrated near the higher end, although slightly lower than those of SMA 3. SMA 5, while initially lagging in the pretest, demonstrated remarkable progress in the posttest, achieving scores that approached those of the other schools. This pattern confirms that the intervention

effectively raised achievement levels across all three schools, even for students with lower initial performance.

Moreover, the narrowing interquartile ranges observed in the boxplot suggest that student performance became more consistent after the intervention, which is particularly notable in SMA 5, where the initial variation was widest. The reduction of variability across schools aligns with the homogeneity test results, which indicated that posttest variances were balanced despite significant differences in pretest conditions. The clustering of scores at the upper end also points to a ceiling effect, where many students reached or approached the maximum score. From an educational perspective, this indicates not only the success of ChatGPT-assisted learning in improving English learning independence but also its role in equalising outcomes across diverse student groups.

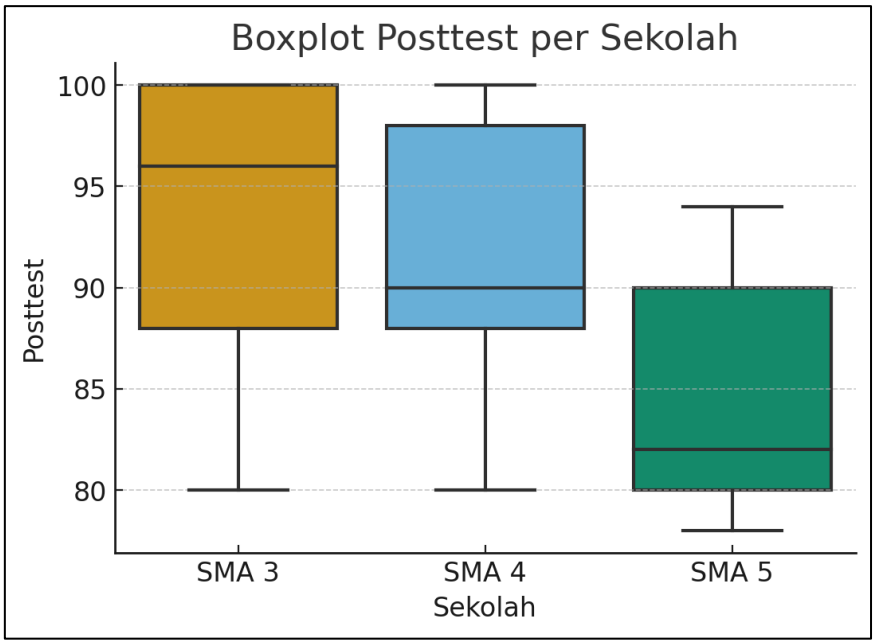

Figure 7: Boxplot of Posttest Scores per School

When comparing the results of **Figures 1** to **6**, apparent differences emerge between the pretest and posttest distributions for the three schools (SMA 3, SMA 4, and SMA 5). For SMA 3, the pretest scores (Figure 1) exhibited a distribution that closely aligned with the theoretical normal distribution, indicating approximate normality. However, the posttest scores (Figure 2) displayed substantial deviations at the upper tail, with many students clustering near the maximum score. This suggests a ceiling effect and confirms the Shapiro-Wilk test result that posttest data in SMA 3 were not normally distributed. For SMA 4, the pretest distribution (**Figure 3**) already showed deviations from normality, with points in both the lower and upper tails diverging from the line. This indicates that students' baseline performance was not normally distributed, which was supported by the Shapiro-Wilk result ($p = 0.035$). In the posttest distribution (**Figure 4**), the deviation became more apparent at the upper tail, with many students clustering at high scores, reflecting substantial performance improvement. The posttest scores,

therefore, not only remained non-normal but also highlighted a ceiling effect similar to SMA 3, confirming the statistical test ($p = 0.008$).

In SMA 5, the pretest distribution (**Figure 5**) exhibited significant deviations from the normal line, particularly in the lower tail, suggesting that student baseline performance was highly variable and not normally distributed. This corresponds with the Shapiro-Wilk test result ($p = 0.021$). However, after the intervention, the posttest scores (**Figure 6**) showed a more consistent distribution, though deviations at the tails remained. Like SMA 3 and SMA 4, many students clustered at higher scores, reducing variability but also creating non-normality, consistent with the Shapiro-Wilk result ($p = 0.001$). Overall, the comparison across all six figures reveals a similar pattern: the pretest data in SMA 3 were closer to normality compared to SMA 4 and SMA 5.

In contrast, the posttest data across all schools deviated strongly from normality due to clustering at high scores. This trend suggests that the intervention was highly effective in improving student performance; however, it also led to ceiling effects that reduced variability and distorted normality. These patterns align with previous research demonstrating that effective instructional interventions often lead to score compression at the upper end of the distribution, producing ceiling effects that mask individual differences (Fan & Konold, 2010; Johnson & Christensen, 2024). Similar studies in language learning contexts have also reported that targeted instructional support can substantially elevate student performance, resulting in reduced variability and non-normal score distributions (Dörnyei, 2014; Hattie, 2012). From a theoretical standpoint, the observed improvements are consistent with mastery learning principles, which posit that when instruction is effectively differentiated and scaffolded, a large proportion of learners can achieve high levels of competence (Guskey, 2022). Thus, the distributional shifts found in this study support prior evidence that well-designed interventions can produce significant gains across diverse student groups, even when such gains lead to statistical deviations from normality.

Conclusion

The findings indicate that the learning intervention implemented in SMA 3, SMA 4, and SMA 5 was highly effective in improving student performance, as shown by the substantial increase in mean scores between the pretest and posttest, with all paired sample t-tests yielding highly significant results ($p < 0.001$). SMA 3 achieved the highest overall performance, while SMA 5 demonstrated the greatest improvement despite its lower initial scores. Normality tests revealed that only the SMA 3 pretest data met normality assumptions, whereas all other datasets—particularly the posttests—showed clear deviations due to score clustering at the upper end, indicating a ceiling effect. The homogeneity test further showed significant variation in pretest scores across schools but homogeneity in posttest scores, suggesting that the intervention not only enhanced outcomes but also reduced performance disparities. These results confirm the effectiveness of the instructional approach across different school contexts; however, the

widespread violation of normality assumptions underscores the need for non-parametric analytical techniques to ensure more accurate interpretation of the data.

References

AbuSahyon, A. S. E., Alzyoud, A., Alshorman, O., & Al-Absi, B. (2023). AI-driven technology and chatbots as tools for enhancing English language learning in the context of second language acquisition: A review study. International Journal of Membrane Science and Technology, 10(1), 1209–1223.

Benvenuti, M., Cangelosi, A., Weinberger, A., Mazzoni, E., Benassi, M., Barbaresi, M., & Orsoni, M. (2023). Artificial intelligence and human behavioral development: A perspective on new skills and competences acquisition for the educational context. Computers in Human Behavior, 148, 107903.

Chan, R. Y., Sharma, S., & Bista, K. (2024). ChatGPT and global higher education: Using artificial intelligence in teaching and learning. STAR Scholars Press.

Chang, D. H., Lin, M. P.-C., Hajian, S., & Wang, Q. Q. (2023). Educational design principles of using AI chatbot that supports self-regulated learning in education: Goal setting, feedback, and personalization. Sustainability, 15(17), 12921.

Dizon, G., & Tang, D. (2023). Exploring the use of ChatGPT for second language learning. Language Learning & Technology, 3(27), 1–15.

Dörnyei, Z. (2014). Researching complex dynamic systems:'Retrodictive qualitative modelling'in the language classroom. Language Teaching, 47(1), 80–91.

Dwivedi, Y. K.; Kshetri, N.; Hughes, L.; Slade, E. L.; Sharma, S. K.; Tiwari, M. K.; Gautam, A.; Kar, A. K. (2023). The rise of generative AI and its implications for education. International Journal of Information Management, 71, 102642.

Fan, X., & Konold, T. R. (2010). Statistical significance versus effect size.

Field, A. (2024). Discovering statistics using IBM SPSS statistics. Sage publications limited.

Fraenkel, J. R., & Wallen, N. E. (1990). How to design and evaluate research in education. ERIC.

Guskey, T. R. (2022). Implementing mastery learning. Corwin Press.

Hakim, L. (2022). Peranan kecerdasan buatan (artificial intelligence) dalam pendidikan. Kemenristek Dirjen Guru Dan Tenaga Kependidikan, 1.

Hattie, J. (2012). Visible learning for teachers: Maximizing impact on learning. Routledge.

Javaid, M., Haleem, A., Singh, R.P. and Suman, R. (2021). Artificial Intelligence Applications for Industry 4.0: A Literature-Based Study. Journal of Industrial Integration and Management, 7, 83–111. https://doi.org/10.1142/s2424862221300040

Johnson, R. B., & Christensen, L. B. (2024). Educational research: Quantitative, qualitative, and mixed approaches. Sage publications.

Kasneci, E., Seßler, K., Küchemann, S., Bannert, M., Dementieva, D., Fischer, F., Gasser, U., Groh, G., Günnemann, S., & Hüllermeier, E. (2023). ChatGPT for

good? On opportunities and challenges of large language models for education. Learning and Individual Differences, 103, 102274.

Kemendikbudristek. (2022). Keputusan Kepala Badan Standar, Kurikulum, dan Asesmen Pendidikan Nomor 008/H/KR/2022 tentang Capaian Pembelajaran pada Pendidikan Anak Usia Dini, Jenjang Pendidikan Dasar, dan Jenjang Pendidikan Menengah pada Kurikulum Merdeka.

Mohebi, L. (2024). Empowering learners with ChatGPT: Insights from a systematic literature exploration. Discover Education, 3(1), 36.

Mulyasa, H. E. (2023). Implementasi kurikulum merdeka. Bumi Aksara.

Muthohar, S., Filasofa, L. M. K., Azzahra, H. K., & Nabila, A. F. (2025). Artificial Intelligence untuk pendidikan keguruan perspektif mahasiswa internasional dan implikasi untuk Pendidikan Islam. Ta'dibuna: Jurnal Pendidikan Islam, 14(1), 1–24.

Perdana, G. A. (2024). Revolusi cerdas: membuka pintu menuju masa depan pendidikan dengan AI. CV Brimedia Global.

Qassrawi, R. M., ElMashharawi, A., Itmeizeh, M., & Tamimi, M. H. M. (2024). AI-powered applications for improving EFL students' speaking proficiency in higher education. Assessment, 8(9).

Reza, F., Rohmah, Z., & Abdullah, N. N. (2023). Challenges in implementing Kurikulum Merdeka for EFL teachers. JEELS (Journal of English Education and Linguistics Studies), 10(2), 439–469.

Sari, A. R., & Setiawan, B. (2023). Teachers' readiness to integrate artificial intelligence in Indonesian classrooms. Journal of Educational Research and Innovation, 11(112–125).

Sreen, A. H. S., & Majid, M. H. M. (2024). Leveraging ChatGPT for personalized learning: A systematic review in educational settings. Amandemen: Journal of Learning, Teaching and Educational Studies, 2(1), 63–78.

Sukardi, H. M. (2021). Metodologi Penelitian Pendidikan: Kompetensi Dan Praktiknya (Edisi Revisi). Bumi Aksara.

Takona, J. P. (2024). Research design: qualitative, quantitative, and mixed methods approaches. Quality & Quantity, 58(1), 1011–1013.