Vol. 11, No. 2, 2026, pp. 375-386
DOI: https://doi.org/10.29210/30037018000

Contents lists available at Journal ELORA

JRTI (Jurnal Riset Tindakan Indonesia)
ISSN: 2502-079X (Print) ISSN: 2503-1619 (Electronic)

Journal homepage: https://jrti.eloracenter.org/jrti



Does generative-ai feedback in formative assessment improve higher-order thinking: motivational pathways in vocational safety education

Tri Juliantika, Waskito Waskito*), Refdinal Refdinal, Yeka Hendriyani
Department of Technology and Vocational Education, Graduate Program, Faculty of Engineering, Universitas Negeri Padang, Indonesia

Article Info

Article history:

Received Aug 11th, 2026
Revised Sept 22th, 2026
Accepted Oct 17th, 2026

Keyword:

Formative assessment,
Artificial intelligence,
Learning motivation,
Higher-order thinking skills,
Occupational health and safety,
PLS-SEM

ABSTRACT

Occupational health and safety (OHS) education in vocational high schools requires higher-order thinking skills (HOTS), yet formative assessment is often limited by delayed teacher feedback. This study examined whether generative AI-supported formative assessment using ChatGPT and Gemini improved HOTS and learning motivation compared with conventional teacher-led assessment, and explored the relationship between perceptions of AI-assisted assessment, motivation, and HOTS. A posttest-only nonequivalent control group quasi-experimental design involved 60 Grade X Machining Engineering students at SMK Negeri 1 Padang, divided into two intact classes of 30 students each. HOTS was measured using a 20-item reasoned multiple-choice test covering Bloom's cognitive levels C4–C6, while motivation and perceptions of AI-assisted assessment were measured using validated questionnaires. Independent-samples t-tests were used for the primary analysis, followed by exploratory partial least squares structural equation modelling (PLS-SEM) in the experimental group. No significant differences were found between the AI-assisted and conventional assessment groups in HOTS scores (81.73 vs. 80.87, $p = .417$, $d = 0.21$) or motivation (54.03 vs. 54.83, $p = .497$, $d = -0.18$). Within the experimental group, perceptions of AI-assisted assessment were positively associated with motivation ($\beta = 0.743$, $p < .001$), and motivation was positively associated with HOTS ($\beta = 0.476$, $p = .023$). Although the indirect effect was statistically significant ($\beta = 0.354$, $p = .042$), its bias-corrected confidence interval included zero, making the mediation evidence inconclusive. These findings do not demonstrate that AI-assisted formative assessment improves HOTS or motivation beyond conventional teacher feedback. However, the observed motivational relationships warrant further investigation using larger samples, baseline measurements, and longer interventions.

© 2026 The Authors.
This is an open access article under the CC BY-NC-SA license (https://creativecommons.org/licenses/by-nc-sa/4.0)

Corresponding Author:

Waskito,
Universitas Negeri Padang
Email: waskito@ft.unp.ac.id

Introduction

Graduates who come from vocational high schools (Sekolah Menengah Kejuruan, SMK) are not only expected to have basic knowledge of safety procedures, but also be able to hazard a guess on the occurrence of hazards and potential risks as well as formulate the right preventive measures for genuine work situations. Mastery of occupational health and safety (OHS) concepts has been argued to support the workplace performance of

prospective vocational graduates (Widyawati, 2020), and OHS literacy has been shown to predict the practical-work discipline of SMK students (Pratama & Sugiarto, 2025). Taken together, these studies point to an association, rather than an established causal link, between OHS mastery and safe, disciplined practice: students need not only to recall safety rules but also to analyze hazards, evaluate risks, and design preventive measures. Such abilities align with the top tiers of the revised Bloom's taxonomy, i.e. analyzing (C4), evaluating (C5), and creating (C6) (Krathwohl, 2002; Nafiati, 2021), which are usually called HOTS or higher-order thinking skills.

Creating assessment tools that really assess HOTS (Higher Order Thinking Skills) continues to be a difficult task for teachers in practice (Adaby et al., 2026; Sari et al., 2024). Locally, informal interviews with OHS teachers at SMK Negeri 1 Padang indicated that feedback on exercises was still given manually and therefore slowly, and that student participation was low. School records also showed that the average OHS score fell from 73 to 63 and 67 across the 2023/2024–2025/2026 academic years. These scores came from teacher-made end-of-semester tests that were not designed to measure C4–C6 skills and whose validity and reliability were not documented; they therefore cannot be read as evidence of a specific HOTS deficit and are reported only as contextual information. The rationale of this study rests instead on the documented difficulty of assessing HOTS (Adaby et al., 2026; Sari et al., 2024) and on the well-established limits of delayed feedback described below, which together indicate that both the quality of assessment and students' engagement with this subject need to be improved.

The formative assessments can be described as the method of assessments that occur during the learning stage mainly to provide feedback for learners to improve. This method has been known to be a source of improvement in learning achievements since Black and Wiliam (1998) pointed out that these assessments can contribute to learners' progress. In the context of vocational education, giving active feedback through formative assessments has led to a greater student performance (Waskito et al., 2022). Nevertheless, the biggest limitation is the large number of students that teachers have per day so time is extremely limited for them to give individualized feedback in a timely manner.

Formative assessment was chosen as the point of intervention, rather than curriculum redesign or teacher professional development, for three reasons. First, the local problem concerns the speed and frequency of feedback, which is precisely what formative assessment targets; curriculum redesign and teacher training address content and pedagogy but do not by themselves shorten the feedback cycle. Second, feedback is among the most powerful influences on achievement, and its effect depends on whether it addresses the task, the process, or self-regulation (Hattie & Timperley, 2007); good feedback practice also supports the self-regulated learning that analyzing, evaluating, and creating require (Nicol & Macfarlane-Dick, 2006). Third, formative assessment with active feedback has already improved achievement in vocational learning (Waskito et al., 2022) and can be implemented within the existing OHS schedule, whereas curricular and training reforms operate over much longer cycles. The open question is therefore not whether feedback matters, but whether automated AI feedback adds value over good teacher feedback when teacher time is limited.

Gen AI (ChatGPT, Gemini etc.) is the future as it can assist teachers in creating assessments (Hanis & Wahyudin, 2024), deliver automated feedback to students along with the personalized learning features (Hidayati, 2025; Hukom, 2025), change the level of difficulty according to the students' level in the practice exercises (Hanazahra et al., 2025). A literature review on how to apply AI in education listed the merits as well as the demerits of using AI (Garzón et al., 2025). On the other side of the discussion, the impact of AI on learner achievement via blended learning was found at a medium level ($g = 0.5$) which means that the result depends on the method of the research, the duration of the teaching intervention, and the size of the sample (Wu et al., 2025). We should be careful if a student gets the AI answers without being challenged to think critically as the habit of expecting an instant solution might not only demotivate one's effort but also reduce one's ability to reason out effectively (Rahmawati, 2025).

It is no surprise that learning motivation is among the mechanisms most often proposed to explain the potential effects of AI-based feedback. Learning motivation is a very important factor in understanding why students are willing to learn. One of the prominent motivational theories is expectancy-value theory (Eccles & Wigfield, 2002). It states that how much effort and persistence learners spend on a learning activity depend upon what they see as the chances of being successful and how much the activity is valuable to them. Another major theoretical framework is ARCS which identifies four components of a learning motive: attention, relevance, confidence, and satisfaction (Keller, 1987). Several empirical investigations have suggested that, for instance, chatbots provide students with feedback that is very helpful and, therefore, very motivational. One piece of research even showed the same effect, a rise of motivation (Fidan & Gencel, 2022; Yin et al., 2024). It was concluded from the results that such an effect (a motivational change) was a common one for general motivation (Aldahlaki, 2026; Hapsari et al., 2024; Keong, 2025; Larasati et al., 2026; Wahyu et al., 2024). Results of Ji et al. (2024) highlight a very crucial issue: continuous learning motivation through AI-based performance

assessment was mediated by perception, expectation confirmation, satisfaction, and perceived usefulness. No direct impact was found on motivational aspects. As the last link in the chain of learning, motivation is closely related to study outcomes at the vocational level (Cao et al., 2025); however, the degree of correlation depends on the motive being examined (An et al., 2025).

To the best of the authors' knowledge, evidence linking AI-assisted formative assessment, learning motivation, and HOTS outcomes in OHS learning at the SMK level is scarce. Most studies on AI and motivation involve university students or are literature reviews (Hapsari et al., 2024; Keong, 2025), and studies that combine a treatment–control comparison with a model of the underlying mechanism are rare. This study addresses that gap with Grade X Machining Engineering students, using HOTS at levels C4–C6 as the outcome. Its primary question is confirmatory: does AI-assisted formative assessment yield higher HOTS and motivation than conventional teacher-led formative assessment? Its secondary question is exploratory: among students who received AI feedback, do perceptions of that feedback relate to HOTS through motivation? Because the primary comparison yielded null results, the study is reported as a null-result study whose contribution lies in estimating the size of the effect against a strong comparison condition and in proposing a testable hypothesis about the motivational mechanism.

The purpose of this study was to test the following five hypotheses:

H1: Students' perceptions of AI-assisted formative assessment positively affect learning motivation;

H2: Learning motivation positively affects HOTS outcomes in OHS learning;

H3: Perceptions of AI-assisted formative assessment directly affect HOTS outcomes;

H4: Learning motivation mediates the relationship between perceptions of AI-assisted formative assessment and HOTS outcomes; and

H5: HOTS outcomes (primary outcome) and learning motivation (secondary outcome) differ between students who receive AI-assisted formative assessment (experimental group) and students who receive conventional formative assessment (control group).

Method

Research Design

This study used a quantitative approach with a posttest-only nonequivalent control group quasi-experimental design (experimental group: AI-assisted formative assessment → posttest; control group: conventional formative assessment → posttest). The between-group comparison of posttest HOTS and motivation (H5) was the primary, confirmatory analysis. A partial least squares structural equation modeling (PLS-SEM) analysis of the experimental group only (H1–H4) was a secondary, exploratory analysis of the mechanism, because perceptions of AI-assisted assessment could be measured only among students who used AI. No pretest was administered: the students were in their first semester of Grade X, so no OHS report-card grades existed that could serve as a baseline. Because the classes were intact and no baseline measure was available, the two groups cannot be assumed to have been equivalent at the start. The design is therefore treated as a nonequivalent-groups design, the between-group comparison is interpreted as a comparison of posttest levels rather than as an estimate of a causal effect, and selection bias is discussed as a key limitation.

Participants

The population comprised all Grade X Machining Engineering students at SMK Negeri 1 Padang enrolled in the OHS subject. Two intact classes whose curriculum content and OHS schedules matched were selected; this constitutes convenience sampling of intact classes rather than purposive or random sampling. Which class served as the experimental group and which as the control group was determined by school policy, not by random assignment. Each class had 30 students (N = 60), a number fixed by class size rather than by an a priori power calculation. A sensitivity power analysis in G*Power (Faul et al., 2007; two-tailed independent-samples t-test, $\alpha = .05$, power = .80, n = 30 per group) showed that the design could detect only effects of $d \geq 0.74$; detecting a medium effect (d = 0.50) would require 64 students per group. For the PLS-SEM analysis (n = 30), the inverse square root method (Kock & Hadaya, 2018) indicates a minimum of about 28 cases for a path coefficient of 0.476 but about 50 cases for an indirect effect of 0.354, so the mediation estimates are underpowered. The research activities were conducted according to the school's November 2026 timetable.

Treatment

Both groups studied the same OHS (Occupational Health and Safety) content, covering the 5R work culture and work ethics in industry, hazard identification at the workplace, safe work practices, and emergency procedures, with the same learning indicators and time allocation. The intervention consisted of four sessions over two

weeks. Both groups worked on the same ten teacher-developed HOTS exercises; the groups differed only in how feedback on those exercises was produced: its source (generative AI checked by the teacher vs. the teacher alone), its speed (immediate and on request vs. after the group discussion), and its personalization (only the experimental group could request additional items or explanations tailored to their errors).

Both groups did ten exercises together in small groups of three students. Each group talked and solved the exercises without opening the books or phones they had. The only distinction between the groups was the source from which feedback was obtained. The experimental group students relied on ChatGPT or Gemini to verify their answers. To students it was given prompt guide with basic regulations: writing own answers and reasoning first and not asking for answers directly. In addition to the prompt guide, student were provided with 5 prompt templates, for example, asking for the check their reasoning, get the explanation of incorrect answer request for one or more additional items and the evaluation of all 10 answers at a time. AI feedback usually started by indicating if the answer was right or wrong; it identified which part was wrong and explained the correct answer as well as the why if needed. It provided the students a set of templates to make the AI give hints or some directing questions before the student got the answer. Through the process students had to rely on each other for their thoughts on solving the problem, while the AI just gave their thoughts back on what their discussion could be, not the answer itself. The teacher was in charge of checking the AI's answer, and giving corrections whenever a student's understanding was wrong or might lead to a confusion.

In the control group, the teacher gave feedback to the students through a talk and explanation of the right answers, without the involvement of AI. When the study period was over, both groups took the same HOTS posttest and filled out the learning motivation questionnaire.

Instruments

Three main instruments were used. First, the HOTS test (Y) was a 20-item reasoned multiple-choice test at levels C4, C5, and C6. This was a posttest administered to both groups. Each item gave a score of 0 to 4 depending on the chosen answer's correctness and the quality of reasoning (overall maximum = 80), and the score after processing was determined by the formula of (score ÷ 80). x 100. The degree of content validity was determined based on expert judgment and through analyzing using Aiken's V coefficient. [Authors to insert: number of expert judges, range of Aiken's V values, and the internal-consistency reliability coefficient of the HOTS test (e.g., Cronbach's α or KR-20).]

Second, the learning motivation questionnaire (Z) contained 18 four-point Likertscale items based on the ARCS theory and also the intrinsic extrinsic dimensions of motivation, covering six factors (intrinsic motivation, extrinsic motivation, attention, effort, persistence, and engagement) with 3 items each; 12 items had a positive meaning and 6 negative meaning. Members of the control group and the treatment group filled out this survey post-intervention. Third, the instrument on perceived use of AI-assisted formative assessment (X) included 15 fourpoint Likert-scale items grouped under five categories (feedback speed and quality, item personalization, depth of answer evaluation, ability to difficulty level adaptation, and platform usage ease), with 10 positive and 5 negative items. This questionnaire was administered by the experimental group alone after the intervention because it was tapping into a perception of an experience offered only to that specific group, i.e., the AI itself. Before being submitted to analysis unfavorable items were reverse-coded. The construct validity and reliability of both questionnaires were evaluated in the measurement model (outer loadings, Cronbach's α, composite reliability, AVE, and HTMT). In the final model, Cronbach's α was .859 for learning motivation and .861 for perceptions of AI-assisted assessment, and composite reliability was .891 and .904, respectively (Table 5). Because 22 of the 33 items were removed during purification, these coefficients describe the shortened scales, not the original instruments.

Data Analysis

Descriptive statistics, test of normality (Shapiro, Wilk), test of homogeneity (Levene), test for linearity and the data were analyzed using SPSS 26. The hypothesis H5 was tested with an independent-samples t test which was accompanied by Cohen's d (Cohen, 1992) and 95% confidence interval for mean difference. Because the HOTS scores of the experimental group were not normally distributed, a Mann–Whitney U test was run as a sensitivity check, and the post hoc power of each comparison was calculated. Because two outcomes were compared, a Bonferroni-adjusted α of .025 was also applied. Hypotheses H1–H4 were tested in a secondary, exploratory analysis using partial least squares structural equation modeling (PLS-SEM) with the program SmartPLS version 4. Here the researchers only used data from the experimental group (n = 30) since variable X was only available to that group. PLS-SEM with path weighting schemes that is based on iterations (maximum was 300). The criterion for stopping was 10^{-7}. The measurement model was analyzed on outer loadings, Cronbach's alpha, composite reliability, AVE, and HTMT with the loading in the final model being ≥ 0.60, AVE ≥ 0.50, and HTMT < 0.90 as thresholds. All variables that showed less than 0.60 of loading and that caused the level of AVE to drop short of the desired threshold, along with those variables that gave rise to HTMT greater than the

limit, were removed in a successive (stepby-step) removal (item purification) method HOTS outcomes (Y) were used as a single-indicator construct. The structural model was analyzed using R^2, f^2, VIF, and SRMR/NFI as secondary indicators (Hair et al., 2019). To test the significance of paths bootstrapping method of 5,000 subsamples (95%) was used and no sign changes were allowed. The significant paths were those with t value > 1.96 and also those that showed p-values below.05. Mediation type was established on the basis of variance accounted for (VAF): mediation was said to be full if VAF was > 80%, partial if it was between 20 and 80%, and no mediation if the VAF was less than 20% (Hair et al., 2017). Because the path tests were exploratory, they were not corrected for multiple testing; mediation was judged primarily by whether the bias-corrected bootstrap interval of the indirect effect excluded zero, and VAF was used only to describe its magnitude.

Results and Discussion

Descriptive Statistics

Table 1 reports the descriptive statistics. The average highorder thinking score of the experimental group was 81.73, which was a little higher than that of the control group, 80.87. On the other hand, the average learning motivation score of the control group was 54.83, slightly higher than the experimental group, which was 54.03. The AI-assisted assessment perception scores in the experimental group were between 36 to 53 (M = 44.23; SD = 3.77) with the total being 60. The HOTS score range was somewhat narrow as it was the case for all students (74–89). No student reached the maximum score: the highest score was 89, 11 points below the ceiling of 100, so a strict ceiling effect was not observed. However, the restricted range (SD ≈ 4 points; coefficient of variation ≈ 5%) limits the test's ability to discriminate between students and reduces the power to detect group differences.

Table 1. Descriptive Statistics of the Research Variables

Variable	Group	n	Min	Max	M	SD
Perception of AI-assisted assessment (X)	Experimental	30	36	53	44.23	3.766
Learning motivation (Z)	Experimental	30	44	63	54.03	4.375
Learning motivation (Z)	Control	30	47	64	54.83	4.676
HOTS outcome (Y)	Experimental	30	75	88	81.73	4.315
HOTS outcome (Y)	Control	30	74	89	80.87	3.884

Note. Scores for X (15 items) and Z (18 items) are summed item scores on a 1–4 scale after reverse-scoring unfavorable items; Y is scored on a 0–100 scale.

Assumption Tests

A total of four out of the five data groups satisfied the normality condition that the data were not skewed. The experimental group's HOTS results had a significant departure from the normal distribution as W turned out to be 0.919 with p being.026; this non-normality stemmed from scores that were packed between 80-81 (8 students out of 30) not from the influence of extreme values or outliers. Giving consideration to the fact that both groups were comparable in size, as well as the the fact that the t-test has been found to work well even when normality is violated slightly, thet-test was not abandoned. Homogeneity of variances was the case and also the relationships between variables were linear. As a sensitivity check for the non-normal HOTS distribution, the Mann–Whitney U test [authors to insert: U, z, p, and effect size r] was compared with the t-test result [authors to state whether the conclusion of no significant difference is unchanged].

Table 2. Results of the Assumption Tests

Test	Variable (group)	Statistic	p	Decision
Shapiro–Wilk	AI perception, X (experimental)	0.952	.197	Normal
Shapiro–Wilk	Motivation, Z (experimental)	0.983	.891	Normal
Shapiro–Wilk	Motivation, Z (control)	0.934	.061	Normal
Shapiro–Wilk	HOTS outcome, Y (experimental)	0.919	.026	Not normal
Shapiro–Wilk	HOTS outcome, Y (control)	0.971	.576	Normal
Levene	HOTS outcome, Y	F = 1.086	.302	Homogeneous
Levene	Motivation, Z	F = 0.108	.744	Homogeneous
Linearity	X → Y	–	.160	Linear
Linearity	X → Z	–	.429	Linear
Linearity	Z → Y	–	.094	Linear

Note. For the linearity tests, p refers to the significance of deviation from linearity.

Between-Group Differences (H5)

The independent-samples t-test showed no significant differences between the experimental and control groups in HOTS outcomes or learning motivation (Table 3). The HOTS difference was 0.86 points (d = 0.21, small), and its 95% confidence interval (−1.26 to 2.98) included zero, so it is compatible with both a small negative and a moderate positive effect. For motivation, the difference was −0.80 points (d = −0.18; 95% CI −3.14 to 1.54).

Both results remained non-significant after Bonferroni correction (α = .025). The post hoc power was only .13 for the HOTS difference and .11 for the motivation difference; with 30 students per group, a two-tailed test with α = .05 and power of .80 could detect only effects of d ≈ 0.74 or larger, and detecting an effect of d = 0.21 would require about 357 students per group. Hence, the absence of significance cannot be taken as evidence that small or moderate effects are absent. H5 was therefore not supported.

Table 3. Results of the Independent-Samples t-Test

Variable	M (SD) experimental	M (SD) control	Difference [95% CI]	t(58)	p	d
HOTS outcome (Y)	81.73 (4.315)	80.87 (3.884)	0.86 [−1.26, 2.98]	0.818	.417	0.21
Learning motivation (Z)	54.03 (4.375)	54.83 (4.676)	−0.80 [−3.14, 1.54]	−0.684	.497	−0.18

Note. The equal-variance assumption was met. Values of t, p, and d were obtained from the SPSS output; the degrees of freedom (58) and confidence intervals were computed from the summary statistics in Table 1.

Measurement Model

The original model comprised 15 X and 18 Z items. Through stepwise assessment, 22 out of the 33 items were removed (Table 4), 12 being the ones whose loadings were less than 0.40 (4 being negative although they had been appropriately reverse-scored), 8 items that were dropped so that AVE would be above 0.50 (initially AVE was roughly 0.35 for X and 0.44 for Z), and 2 because the HTMT between X and Z (0.967) was over 0.90. The resulting model consisted of four X items and seven Z items. The surviving Z items were spread over all six motivation dimensions while for the X items' case, only three of the five dimensions were still there, namely quality of feedback explanations, depth of answer analysis, and ease of use; the dimension of item personalized and difficulty level adap- tivity were entirely missing in this version.

Table 4. Summary of the Item Elimination Process

Stage	Items eliminated	Reason
Low or negative loading (< 0.40)	X3, X9, X10, X11, X15; Z1, Z3, Z6, Z9, Z12, Z15, Z18 (12 items)	Loadings of X15, Z6, Z15, and Z18 were negative despite reverse-scoring
Convergent validity (AVE ≥ 0.50)	X1, X5, X6, X7, X12; Z5, Z8, Z13 (8 items)	Loadings of 0.40–0.70; eliminated from the lowest until AVE ≥ 0.50
Discriminant validity (HTMT < 0.90)	X4, Z17 (2 items)	X4 loaded higher on Z (0.573) than on X (0.447); Z17 loaded almost equally on both constructs; HTMT decreased from 0.967 to 0.792
Final model	X2, X8, X13, X14; Z2, Z4, Z7, Z10, Z11, Z14, Z16	Met the validity and reliability criteria

All loadings of variables in the final model were > 0.632, both constructs also satisfied the reliability and convergent validity standards (Table 5). Discriminant validity was confirmed in all pairwise combinations (HTMT: X, Z = 0.792; X, Y = 0.367; Z, Y = 0.496). Moreover, all variation-inflation factor (VIF) values were < 3.3 (outer model: 1.668–2.567; inner model: 1.000–2.230) which showed that the collinearity assumption remained satisfied.

Table 5. Final Measurement Model

Construct	Outer loadings	α	CR (rho_c)	AVE	Items
Perception of AI-assisted assessment (X)	X2 = 0.884; X8 = 0.883; X13 = 0.824; X14 = 0.758	0.861	0.904	0.704	4
Learning motivation (Z)	Z2 = 0.773; Z4 = 0.735; Z7 = 0.745; Z10 = 0.811; Z11 = 0.654; Z14 = 0.775; Z16 = 0.632	0.859	0.891	0.540	7
HOTS outcome (Y)	Single indicator = 1.000	–	–	–	1

Note. CR = composite reliability. Composite reliability (rho_a): X = 0.894; Z = 0.872.

Structural Model and Testing of H1–H4

Perceptions of AI-assisted assessment the explanation of learning motivation variance from perceptions of AI-assisted assessment Perceptions of AI-assisted assessment explained 55.1% of the variation in learning motivation (R^2 = 0.551; adjusted R^2 = 0.535; moderate), whereas motivation and perception together could explain only 23.6% of the variation in HOTS outcomes (R^2 = 0.236; adjusted R^2 = 0.180; weak). Model fit indices did not meet the desired thresholds (SRMR = 0.117 > 0.08; NFI = 0.601). In the context of prediction-oriented PLS-SEM, it is true that these indices are auxiliary, however, they have to be viewed as limitations.

Bootstrapping results are summarized in Table 6 and Figure 1. AI-assisted assessment perceptions were positively and greatly linked to learning motivation (H1; β = 0.743; f² = 1.230). Learning motivation was positively connected with HOTS outcomes (H2; β = 0.476; f² = 0.133). However, the bias-corrected confidence interval at this path includes zero. The direct path from perception to HOTS outcomes was approximately zero and non-significant (H3; β = 0.013; f² = 0.000). Indirect effect (X → Z → Y) was significant by the p value alone (H4; β = 0.354; p =.042) which suggests full mediation, with a VAF of 96.5% that exceeds the 80% minimum standard. Unfortunately, the percentile interval of the effect of mediation remained in a positive range (from 0.058 to 0.752) whereas the bias-corrected interval contained zero (−0.080 to 0.644), which indicates that the conclusion depends on the estimation method. So we should make very careful interpretation of our findings.

Table 6. Results of the Structural Model Test (Bootstrapping with 5,000 Subsamples, n = 30)

Hyp.	Path	β	t	p	95% CI	f²	Decision
H1	X → Z	0.743	9.804	< .001	[0.524, 0.849]	1.230	Supported
H2	Z → Y	0.476	2.278	.023	[−0.089, 0.801]	0.133	Partially supported
H3	X → Y (direct)	0.013	0.064	.949	[−0.377, 0.442]	0.000	Not supported
H4	X → Z → Y (indirect)	0.354	2.037	.042	Percentile [0.058, 0.752]; BC [−0.080, 0.644]	–	Partially supported

Note. CI = confidence interval; BC = bias-corrected. Intervals for H1–H3 are bias-corrected. "Partially supported" indicates p < .05 but a bias-corrected interval that includes zero. H5 was tested separately (Table 3) and was not supported.

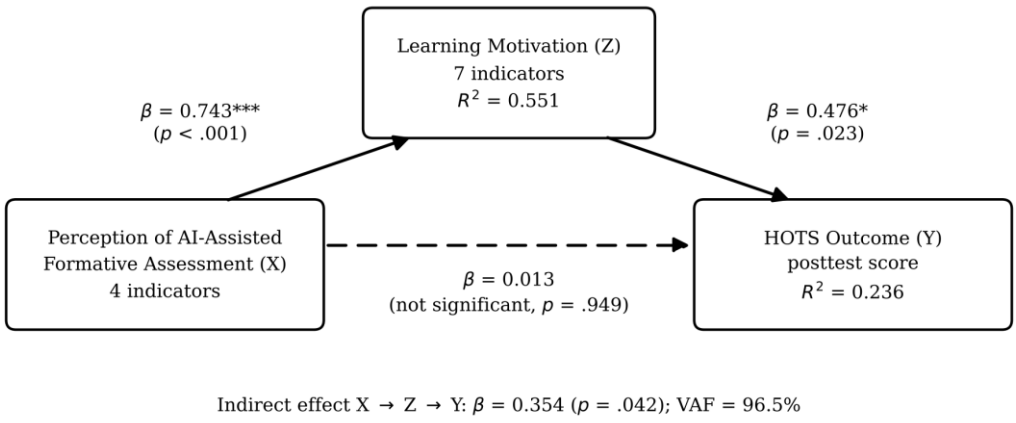


Figure 1. Final Structural Model and Path Coefficients

*Note. Final bootstrapped model for the experimental group (n = 30). The dashed line indicates a non-significant path; *** p < .001; * p < .05. The bias-corrected intervals for the Z → Y path and the indirect effect include zero (see Table 6).*

Absence of Between-Group Differences in HOTS Outcomes and Motivation

The main takeaway is that AI-assisted formative assessment does not generate significantly different results regarding high order thinking skills (HOTS) and level of motivation in contrast with traditional formative assessment methods. The results were in favor of the experimental group for HOTS outcomes (d = 0.21), but the difference turned out to be very low since confidence interval had 0 as well. Meanwhile, motivation was a bit less in the experimental group (d = −0.18). The effect observed here is much less pronounced than a moderate effect of AI on learning achievement reported in a meta-analysis (Wu et al., 2025). Many reasonable explanations that can't be figured out with the available data need to be explored.

Wu et al. (2025) reported that the magnitude of AI effects depends on intervention duration, design, and sample size. In this study, 30 students per group was considered a small sample size while not being sure that the treatment duration was enough. Therefore, it was most responsive to fairly large effects only. We also considered a strong comparison condition, and for that reason, the control group was not a no-assessment one but it kept doing HOTS practice, teacher checking, and feedback. It is a kind of practice that has already been found to lead to improved achievements of vocational students (Waskito et al., 2022). The distance between a feedback intervention and another meaningful form of feedback intervention could be much smaller than the distance between a no-feedback condition and a feedback condition. Measurement aspects HOTS scores were rather limited as only in the range of 74–89 (mean 100) with only 80-odd out of 100, which might lead to such a problem as the result cannot reach the maximum (a ceiling effect), or the test is weak at differentiating at the

higher levels (C4, C6). This explanation needs to be verified using item analysis. There was no pretest conducted; that is, no baseline measurement was performed, a situation which makes it impossible to distinguish between minor variation at the end of a treatment period and the initial, inherent differences between the two groups of classes.

In addition to the main idea of how AI was operationalized, the following method of AI application should also be kept in mind. In the study mentioned here, AI was used after students had been given a chance to discuss, answer, and check the items together. The students were even given AI-generated explanations to those items across all groups. Such a design, which is feedback principles-based where students have to try a task first and receive feedback (Hattie & Timperley, 2007; Roediger & Karpicke, 2006), is in fact a good one. However, it also means that the features commonly seen as advantages of AI, i.e., item personalization and adaptive difficulty adjustment, were only a little visible (e.g., only when students asked for additional items). Thus, the findings cannot be transferred/attributed to AI implementation that offers individual-student adaptive practice as the teacher was also watching the AI feedback and correcting misconceptions, meaning AI here was more a support tool than a total replacement. A setup like the present one is indeed in accordance with the worry that AI in the absence of criticalthinking activity will cause the decline of students' interest and capability to analyse (Rahmawati, 2025), and hence, it can be inferred that the effects seen here are the results of the "AI plus teacher supervision" package, not "AI alone".

Perceptions of AI-Assisted Assessment and Learning Motivation (H1)

Within the experimental group, students who evaluated AI-assisted assessment more positively tended to report higher learning motivation ($\beta = 0.743$; $f^2 = 1.230$). This direction is consistent with expectancy–value theory (Eccles & Wigfield, 2002) and the ARCS model (Keller, 1987), which link immediate and specific feedback with confidence and satisfaction, as well as with studies reporting a relationship between AI-based feedback and motivation (Fidan & Gencel, 2022; Larasati et al., 2026; Yin et al., 2024). This finding also resembles the pattern reported by Ji et al. (2024), who found that AI-based performance assessment was related to motivation through students' perceptions and satisfaction.

In other words, the size of the coefficient should not be taken as evidence of treatment effect. First, the comparison of the two groups indicated no difference regarding motivation; as a result, the relationship observed within the experimental group had not resulted in group-level advantage. Secondly, X (perception) and Z (motivation) have been assessed using the same questionnaire, at the same time, and by the same respondents; motivated students may have perceived their learning as more pleasant (reverse causality), or both measures have been impacted by similar response tendencies (common method variance). Third, the high correlation matched with the empirical overlap between the two constructs in the initial model (HTMT = 0.967) and the construct X was made of just 4 items which means not measuring the dimensions of personalization and adaptivity that are the main advantages of AI. Thus, one could conclude that the perception assessed by the final measure is more properly described as understanding and ease-of-use of AI rather than perception of adaptive evaluation as a whole.

Learning Motivation and HOTS Outcomes (H2)

Learning motivation was the strongest predictor of HOTS outcomes at around 0.476 with statistical significance at 0.023. The findings support earlier results on vocational education revealing both intrinsic and extrinsic motivation have an influence on academic achievement (Cao et al., 2025). However, the confidence interval calculated using the bias-corrected method is quite large, ranging from −0.089 to 0.801; in other words, the true effect could range from negligible to strong. The R squared for HOTS outcomes was low only 0.236 and also that the path was not significant in the previous model without item purification. The model estimates can vary a lot even when the sample size is as small as 30. Furthermore, the motivation-achievement relationship identified in these two studies seems rather narrow as only certain types of learning motivation were considered in An et al. (2025). Such variation in types of motivation is likely due to the fact that different types of motivation can have vastly different effects on achievement and that motivation is often measured in many ways. In this paper, motivation was modeled as a single reflective construct so dimensional differences are not taken into account.

Direct Path and Mediation (H3 and H4)

There was little or no direct path from perception to Higher Order Thinking Skills (HOTS) outcomes ($\beta = 0.013$) and the mediated pathway through motivation showed a very large variation (96.5%) which fully conforms to the definition of full-mediation. Another "no directeffect/only indirect effect" model was found by Ji et al. (2024) with AI and learning motivation, and a similar finding has been reported where the role of technology in influencing student achievement is mostly indirect or via other intermediary factors (Jian, 2022; Simões et al., 2022). However, there are four reasons to believe the mediation conclusion is still tentative. (1) The indirect effect has different confidence intervals: one was at zero and another excluded it. (2) Three variables were only

measured once at the same time as the post-treatment; hence, the causal chain X → Z → Y which mediation inference demands could not be guaranteed. (3) X is the perception that is only available to the experimental group; therefore, the structural model did not make use of a comparative control group. (4) The t

-test results (H5) did not find any benefit in the experimental group for motivation nor achievement. Therefore, rather than interpreting the findings as evidence that AI-assisted assessment leads to improved HOTS via motivation, it is probably more appropriate to view it as a proposed mechanistic hypothesis that has to be verified experimentally.

Implications

These findings indicate for the teachers of SMK schools that simply introducing AI to formative assessment will not necessarily result in improved HOTS, the fact that may not be the most crucial at all. What really counts are students' perception of the AI feedback as the main element: the clarity, helpfulness, and ease of use, which is complemented by the teachers' vigilance with the quality check of the feedback provided. As regards the theoretical contributions, the present paper puts forth an X → Z → Y model which could be re-validated through the application of designs that permit making stronger inferences as opposed to merely providing confirmatory evidence.

Theoretically, the combination of null between-group results and strong within-group associations points to a boundary condition rather than to a failure of motivational theory. Expectancy–value theory and the ARCS model predict that feedback raises motivation when it increases expectancies of success and perceived value (Eccles & Wigfield, 2002; Keller, 1987). In this study, students who appraised the AI feedback more favorably were more motivated, yet exposure to AI feedback did not raise average motivation. This pattern suggests that what matters motivationally is the individual appraisal of feedback quality, not the delivery channel: when the comparison condition already provides timely, task-focused teacher feedback, AI feedback offers little additional expectancy or value. Feedback theory leads to the same conclusion, because the effect of feedback depends on whether it moves learners beyond task-level correction toward process-level and self-regulatory feedback (Hattie & Timperley, 2007; Nicol & Macfarlane-Dick, 2006), whereas in this implementation AI mainly verified answers, a task-level function the teacher was already performing. The added value of AI in formative assessment may therefore be conditional on (a) a weak or slow baseline feedback condition, (b) AI functions that go beyond verification, such as adaptive items and process-level hints, and (c) sufficient duration for motivational changes to translate into HOTS. These conditions are testable propositions for future research rather than conclusions of the present study.

Limitations

Some restrictions that come with this research are (1) limited number of participants from a single school and the only academic program in question, n = 30 per cohort, which, in turn, reduces the chances for making statistically valid inferences and generalizing the results; (2) assignment of the students to groups is according to intact classes not having a pretest or baseline-ability covariate, therefore the possible threats to internal validity are not addressed; (3) the fact that 22 out of 33 questionnaire items were removed together with huge changes in path coefficients across the stages, e.g., X → Z decreased from 0.934 to 0.743 which makes the PLS SEM results dependent on certain modellings and exclude a confirmatory reading; (4) model fit indices (SRMR = 0.117) do not meet the preferred threshold; (5) reliance on self- reported, psychological measures collected at a single time point; (6) lack of an objective measure of AI-use intensity by students and possible variability of generated content of ChatGPT and Gemini across sessions; (7) the absence of predictive Q^2 indices; and (8) the non-normal distribution and restricted range of HOTS scores. Future studies should therefore use larger multi-school samples with random allocation where feasible, baseline measures as covariates (e.g., ANCOVA), longer interventions with AI-use logs, repeated measurement of motivation to establish temporal order, expanded item pools that retain the personalization and adaptivity dimensions, and qualitative data on students' interactions with AI.

Conclusion

This study compared AI-assisted and conventional formative assessment among 60 Grade X Machining Engineering students at SMK Negeri 1 Padang. AI-assisted formative assessment did not produce significantly higher HOTS ($d = 0.21$; $p = .417$) or motivation ($d = -0.18$; $p = .497$) than conventional teacher-led formative assessment, and these results held after Bonferroni correction ($\alpha = .025$). With a post hoc power of only .13 and a minimum detectable effect of $d \approx 0.74$, however, the data cannot distinguish a true null effect from a small effect that the sample was too small to detect. Theoretically, the findings suggest a boundary condition: when students already receive timely, task-focused feedback from the teacher, AI feedback that mainly verifies answers adds little expectancy or value, and therefore little motivational or cognitive benefit. Within the AI group, more

favorable perceptions of AI-assisted assessment were associated with higher motivation (β = 0.743), and motivation was associated with HOTS (β = 0.476), while the direct path was negligible. This pattern is consistent with motivational mediation, but the indirect effect (p = .042, percentile bootstrap) was not robust: its bias-corrected interval included zero and it would not survive a Bonferroni-adjusted α of .025. Together with simultaneous measurement and substantial item reduction, this makes the mediation finding indicative only. The present data are therefore insufficient to conclude that AI-assisted formative assessment enhances students' higher-order thinking in OHS learning; the motivational pathway is best regarded as a hypothesis to be tested.

Future research should test this hypothesis using the design improvements outlined in the Limitations section.

Acknowledgments

The authors are grateful to Mr. Waskito, who has provided us with the supervision, SMK Negeri 1 Padang, where we had the research facility, also, we thank all Grade X Machining Engineering students who kindly participated in our study.

References

Adaby, M. A., Sugiarto, T., Arif, A., & Muslim. (2026). Analisis tingkat pengetahuan pada soal UAS produktif otomotif semester genap 2024/2025 kelas XII TKR SMK Negeri Teknologi di Kota Padang [Analysis of knowledge levels in productive automotive final-semester examination items for Grade XII light-vehicle engineering students]. *Ensiklopedia of Journal, 8*(2), 98–105. https://doi.org/10.33559/eoj.v8i2.3710

Aldahlaki, Z. H. A. (2026). The role of artificial intelligence in enhancing learning motivation through self and immediate feedback. *Journal of Technology and System Information, 3*(1), 1–7. https://doi.org/10.47134/jtsi.v3i1.5570

An, P. M. N., Anh, T. N. P., & Dinh, C.-T. (2025). Investigating the mediating effects of motivation on academic performance in a synchronous EFL online learning. *Journal of e-Learning and Knowledge Society, 21*(1), 224–235. https://doi.org/10.20368/1971-8829/1135976

Black, P., & Wiliam, D. (1998). Assessment and classroom learning. *Assessment in Education: Principles, Policy & Practice, 5*(1), 7–74. https://doi.org/10.1080/0969595980050102

Cao, C., Taddese, E. T., Gebresilase, B. M., Shorouk, A., & Berhanu, K. Z. (2025). The influence of learning motivation on academic performance in Chinese vocational college students: A self-determination theory perspective. *International Journal of Innovative Research and Scientific Studies, 8*(6), 3297–3309. https://doi.org/10.53894/ijirss.v8i6.10339

Cohen, J. (1992). A power primer. *Psychological Bulletin, 112*(1), 155–159. https://doi.org/10.1037/0033-2909.112.1.155

Eccles, J. S., & Wigfield, A. (2002). Motivational beliefs, values, and goals. *Annual Review of Psychology, 53*, 109–132. https://doi.org/10.1146/annurev.psych.53.100901.135153

Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods, 39*(2), 175–191. https://doi.org/10.3758/BF03193146

Fidan, M., & Gencel, N. (2022). Supporting the instructional videos with chatbot and peer feedback mechanisms in online learning: The effects on learning performance and intrinsic motivation. *Journal of Educational Computing Research, 60*(7), 1716–1741. https://doi.org/10.1177/07356331221077901

Garzón, J., Patiño, E., & Marulanda, C. (2025). Systematic review of artificial intelligence in education: Trends, benefits, and challenges. *Multimodal Technologies and Interaction, 9*(8), 84. https://doi.org/10.3390/mti9080084

Hair, J. F., Hult, G. T. M., Ringle, C. M., & Sarstedt, M. (2017). *A primer on partial least squares structural equation modeling (PLS-SEM)* (2nd ed.). SAGE Publications.

Hair, J. F., Risher, J. J., Sarstedt, M., & Ringle, C. M. (2019). When to use and how to report the results of PLS-SEM. *European Business Review, 31*(1), 2–24. https://doi.org/10.1108/EBR-11-2018-0203

Hanazahra, A., Efriyanti, L., Satria, E., & Indra, Y. (2025). Evaluasi implementasi pembelajaran adaptif berbasis artificial intelligence (AI) di SMKN 1 Bukittinggi [Evaluation of the implementation of artificial intelligence-based adaptive learning at SMKN 1 Bukittinggi]. *Jurnal Pendidikan Teknologi Informasi dan Vokasional, 7*(2), 1–13. https://doi.org/http://dx.doi.org/10.23960/jpvti

Hanis, M., & Wahyudin, D. (2024). Pemanfaatan artificial intelligence (AI) dalam penyusunan asesmen pembelajaran bagi guru sekolah dasar [Utilizing artificial intelligence (AI) in developing learning

assessments for elementary school teachers]. *Jurnal Ilmiah Profesi Pendidikan, 9*(2). https://doi.org/10.29303/jipp.v9i2.2252

Hapsari, D. D., Ramadhani, G. Y., & Ikramullah, N. I. (2024). Literature review: Pengaruh artificial intelligence (AI) terhadap motivasi belajar peserta didik [Literature review: The effect of artificial intelligence (AI) on students' learning motivation]. *Jurnal Empati, 13*(4), 313–324. https://doi.org/10.14710/empati.2024.46697

Hattie, J., & Timperley, H. (2007). The power of feedback. *Review of Educational Research, 77*(1), 81–112. https://doi.org/10.3102/003465430298487

Hidayati, H. N. (2025). Pemanfaatan artificial intelligence pada penilaian pembelajaran mendalam [Utilizing artificial intelligence in deep learning assessment]. *Afeksi: Jurnal Penelitian dan Evaluasi Pendidikan, 6*(5), 900–908. https://doi.org/10.59698/afeksi.v6i5.531

Hukom, J. (2025). Model integrasi feedback digital berbasis AI untuk pembelajaran berorientasi kompetensi [An AI-based digital feedback integration model for competency-oriented learning]. *IDENTIK: Jurnal Ilmu Ekonomi, Pendidikan dan Teknik, 2*(4), 169–174.

Ji, H., Suo, L., & Chen, H. (2024). AI performance assessment in blended learning: Mechanisms and effects on students' continuous learning motivation. *Frontiers in Psychology, 15*, 1447680. https://doi.org/10.3389/fpsyg.2024.1447680

Jian, Z. (2022). Sustainable engagement and academic achievement under impact of academic self-efficacy through mediation of learning agility: Evidence from music education students. *Frontiers in Psychology, 13*, 899706. https://doi.org/10.3389/fpsyg.2022.899706

Keller, J. M. (1987). Development and use of the ARCS model of instructional design. *Journal of Instructional Development, 10*(3), 2–10. https://doi.org/10.1007/BF02905780

Keong, L. M. (2025). AI-personalised learning in higher education: A study on learning outcomes and motivation among university students. *International Journal of Research and Innovation in Social Science, 9*(5), 6075–6085. https://doi.org/10.47772/IJRISS.2025.905000473

Kock, N., & Hadaya, P. (2018). Minimum sample size estimation in PLS-SEM: The inverse square root and gamma-exponential methods. *Information Systems Journal, 28*(1), 227–261. https://doi.org/10.1111/isj.12131

Krathwohl, D. R. (2002). A revision of Bloom's taxonomy: An overview. *Theory Into Practice, 41*(4), 212–218. https://doi.org/10.1207/s15430421tip4104_2

Larasati, S. P., Sukma, D., Karneli, Y., Fikri, M., Amat, M. A. C., & Rismi, R. (2026). AI-driven feedback and its impact on student motivation in online learning. *G-COUNS: Jurnal Bimbingan dan Konseling, 10*(2), 1071–1088. https://doi.org/10.31316/g-couns.v10i02.8584

Nafiati, D. A. (2021). Revisi taksonomi Bloom: Kognitif, afektif, dan psikomotorik [Revision of Bloom's taxonomy: Cognitive, affective, and psychomotor]. *Humanika: Kajian Ilmiah Mata Kuliah Umum, 21*(2), 151–172. https://doi.org/10.21831/hum.v21i2.29252

Nicol, D. J., & Macfarlane-Dick, D. (2006). Formative assessment and self-regulated learning: A model and seven principles of good feedback practice. *Studies in Higher Education, 31*(2), 199–218. https://doi.org/10.1080/03075070600572090

Pratama, D. H., & Sugiarto, T. (2025). Literasi keselamatan dan kesehatan kerja sebagai prediktor kedisiplinan praktikum siswa SMK program Teknik Kendaraan Ringan [Occupational health and safety literacy as a predictor of practical-work discipline among light-vehicle engineering students]. *JTPVI: Jurnal Teknologi dan Pendidikan Vokasi Indonesia, 3*(4), 891–902. https://doi.org/10.24036/jtpvi.v3i4.222

Rahmawati, Y. (2025). Kecerdasan buatan di ruang kelas: Analisis penurunan motivasi belajar dan kemampuan berpikir kritis siswa SMP Negeri 1 Ganding Sumenep [Artificial intelligence in the classroom: An analysis of declining learning motivation and critical-thinking ability among students of SMP Negeri 1 Ganding Sumenep]. *Jurnal Teknologi Pendidikan, 2*(4), 1–8. https://doi.org/10.47134/jtp.v2i4.1640

Roediger, H. L., III, & Karpicke, J. D. (2006). Test-enhanced learning: Taking memory tests improves long-term retention. *Psychological Science, 17*(3), 249–255. https://doi.org/10.1111/j.1467-9280.2006.01693.x

Sari, I. P., Hatta, T. A., Manurung, A. A., & Dachi, S. W. (2024). The influence of a high-order thinking (HOTS)-based learning model on students' midterm question analysis. *Aksaqila International Humanities and Social Sciences Journal, 3*(2), 58–66. https://doi.org/10.30596/aihss.v3i2.505

Simões, S., Oliveira, T., & Nunes, C. (2022). Influence of computers in students' academic achievement. *Heliyon, 8*, e09004. https://doi.org/10.1016/j.heliyon.2022.e09004

Wahyu, R., Pertiwi, L., Kulsum, L. U., & Hanifah, I. A. (2024). Evaluating the impact of artificial intelligence-based learning methods on students' motivation and academic achievement. *International Journal of Post Axial: Futuristic Teaching and Learning, 2*(1), 49–58. https://doi.org/10.59944/postaxial.v2i1.279

Waskito, Irzal, Wulansari, R. E., & Ya, K. Z. (2022). The adventure of formative assessment with active feedback in the vocational learning: The empirical effect for increasing students' achievement. *Journal of Technical Education and Training, 14*(1), 54–62. https://doi.org/10.30880/jtet.2022.14.01.005

Widyawati, N. K. (2020). Pentingnya penguasaan konsep keselamatan dan kesehatan kerja (K3) dalam mendukung kinerja calon lulusan pendidikan kejuruan di dunia kerja [The importance of mastering occupational health and safety (OHS) concepts in supporting the workplace performance of prospective vocational graduates]. *Jurnal Bosaparis: Pendidikan Kesejahteraan Keluarga, 11*(3), 87–93.

Wu, J., Tlili, A., Salha, S., Mizza, D., Saqr, M., López-Pernas, S., & Huang, R. (2025). Unlocking the potential of artificial intelligence in improving learning achievement in blended learning: A meta-analysis. *Frontiers in Psychology, 16*, 1691414. https://doi.org/10.3389/fpsyg.2025.1691414

Yin, J., Goh, T.-T., & Hu, Y. (2024). Using a chatbot to provide formative feedback: A longitudinal study of intrinsic motivation, cognitive load, and learning performance. *IEEE Transactions on Learning Technologies, 17*, 1378–1389. https://doi.org/10.1109/TLT.2024.3364015