https://doi.org/10.31538/alsuna.v9i1.9689

Alsunā Journal of Arabic and English Language

Integrating Problem-Based Learning and Spinning Wheel Media to Improve Arabic Grammar Mastery in Islamic boarding schools: A Quasi-Experimental Study

Muhammad Fadhli Alkhoiri [1*], Asrina [1], Hanomi[1], Fathi Hisyam Panagara[2]

[1]Universitas Islam Negeri Imam Bonjol, Indonesia

[2]Pusat Studi Pembelajaran Bahasa Arab, Indonesia

Received, 20 January 2026 **Accepted, 02 June 2026**

ABSTRACT

Purpose *- This study examines how Problem-Based Learning (PBL) combined with Spinning Wheel media operates in classical Arabic grammar (nahwu) instruction, and whether this combination is associated with improved qawāʿid nahwu mastery among ninth-grade students in an Islamic boarding school.*

Design/methodology/approach *- A quasi-experimental pretest–posttest nonequivalent control-group design was conducted with 50 ninth-grade students at Al-Hikmah Islamic Boarding School, West Pasaman. Class IX-1 (n = 25, female) was the experimental group; Class IX-2 (n = 25, male) the control. The gender-based class split is a school-level structural constraint and is treated as a major internal validity limitation. Instruments were expert-validated (CVR ≥ .62), pilot-tested (Cronbach's α = .84), and short-answer items were scored with an analytic rubric verified for inter-rater reliability (κ = .82). Statistical analysis included Shapiro–Wilk, Levene's test, independent-samples t-test, and ANCOVA with homogeneity-of-regression-slopes verification. The treatment ran across eight 70-minute sessions over four weeks.*

Findings/results *– Pretest means were statistically comparable ($t(48) = 1.73$, $p = .091$). Posttest comparison yielded $t(48) = 8.14$, $p < .001$, Cohen's $d = 2.01$; ANCOVA $F(1, 47) = 66.3$, $p < .001$, partial $\eta^2 = .585$; N-Gain = .61 (medium–high) versus .17 (low) for the control. Qualitative themes corroborated higher engagement and metalanguage use in the experimental group. These results suggest a strong association between the model and improved mastery; causal attribution remains limited by the quasi-experimental design.*

Originality/value *- This study offers preliminary evidence that embedding the Spinning Wheel as a randomised problem-trigger within PBL's initiation phase may lower affective and cognitive barriers to grammar engagement in classical Arabic contexts — a pedagogical pairing not previously examined under quasi-experimental classroom conditions in Islamic boarding schools.*

Paper type *– Research paper*

Keywords**:** *Problem-Based Learning; Spinning wheel; Arabic grammar; Islamic boarding school.*

***Corresponding author:** Muhammad Fadhli Alkhoiri, Universitas Islam Negeri Imam Bonjol Padang, Indonesia, E-mail: mfadlialkhoiri@gmail.com

1. Introduction

Arabic occupies a central place in Islamic education. Mastery of its grammar is a scholarly requirement: the primary sources of Islamic jurisprudence, exegesis, and creed are written in Arabic, and most classical commentaries remain inaccessible without grammatical

competence. Seff (2019) documented how Arabic instruction in Indonesia has moved from a purely liturgical function toward an academic discipline engaged across research and knowledge production — a shift that places new pressure on how the language, and especially its grammar, is taught.

Nahwu (Arabic morphosyntax) remains the most persistently reported obstacle for students in Islamic boarding schools. Learners struggle with abstract grammatical rules and their application to classical texts. Wahyudin (2025) traced this to practices built around direct explanation and memorisation. Fatahillah and Wahyudin (2025) connected those same practices to declining classroom interaction and the anxiety that stops students from attempting grammatical analysis. The broader SLA literature supports this diagnosis: Cook (2016) and Ellis (2006) both showed that form-focused instruction without problem-solving or communicative tasks suppresses retention and transfer.

Problem-Based Learning (PBL) places an ill-structured problem before learners at the outset, requiring collaborative analysis and solution-building rather than passive reception (Barrows & Tamblyn, 1980; Hmelo-Silver, 2004). Al-Drees et al.'s (2024) systematic review of 27 Scopus- and WoS-indexed studies confirmed that PBL improves academic performance and critical thinking in language-learning contexts. In Arabic and Islamic educational settings, Fauzi (2020) documented significant nahwu gains under PBL; Subhi et al. (2026) argued PBL is particularly suited to language learning because it develops the higher-order skills that grammar acquisition depends on.

The model's effectiveness depends on how the initial problem reaches students. A trigger that draws learners into the problem space before analysis begins matters: game-based and randomised media have shown particular efficacy here. Chen et al.'s (2023) meta-analysis across 41 studies reported g = 0.82 for gamification on learning outcomes, with motivation and engagement as the most consistent gains. The Spinning Wheel serves this role: a digital randomisation tool presenting grammatical problems in unpredictable order, disrupting the anticipatory passivity common in rote-learning environments (Huda, 2020). Ahzan (2025) found it significantly improved comprehension and interaction in Arabic instruction; Huda (2020) was first to apply it in nahwu specifically, finding it effective at converting abstract case-marking problems into discussable situations.

This study embeds the Spinning Wheel within PBL's problem-initiation phase. Each session opens with a wheel rotation selecting a classical Arabic sentence with a grammatical error or parsing ambiguity (iʿrāb); PBL then provides the structured framework for resolving it. Adiana (2024) found this combination enhanced critical thinking beyond either tool in isolation; Akzam and Hayati (2026) called for precisely this integration of traditional content with contemporary pedagogy in Arabic language education.

The problem driving the study is persistent underachievement and low engagement in nahwu instruction at Al-Hikmah Islamic Boarding School, West Pasaman, where students consistently score below the minimum mastery criterion (KKM = 70) under conventional teaching. Two research questions guide the inquiry:

RQ1: How is PBL integrated with Spinning Wheel media applied in nahwu instruction at Al-Hikmah Islamic Boarding School?
RQ2: To what extent is the PBL–Spinning Wheel model associated with improved student mastery of qawāʿid nahwu, as indicated by pretest–posttest comparison, ANCOVA, and normalised gain?

2. Literature Review

2.1 Problem-Based Learning: Theoretical Basis and Language-Learning Evidence

PBL originated in medical education (Barrows & Tamblyn, 1980)and has since been applied across language and humanities instruction (Hmelo-Silver, 2004; Savery, 2006). Its premise is that ill-structured problems, placed before formal teaching, activate self-directed inquiry, collaborative deliberation, and knowledge construction that transfers more durably than content delivered through direct instruction. In SLA terms, this maps onto what Ellis (2006) called task-based language teaching: form is encountered inside meaning-bearing tasks rather than isolated exercises.

The international evidence base is solid. Al-Drees et al. (2024) reviewed 27 Scopus- and WoS-indexed studies and confirmed positive effects on academic performance, critical thinking, and engagement in language-learning contexts. Dolmans et al. (2015) reviewed 21 studies and found a consistent positive effect of PBL on deep learning (mean d = .11), strongest in single-session implementations. Fauzi's (2020) observation that PBL can raise cognitive load during problem-presentation for students unaccustomed to self-directed work is a real constraint, and it is precisely what the Spinning Wheel is designed to address.

Cognitive load theory (Sweller, 1988) explains the mechanism. Classical Arabic morphosyntax has high intrinsic load: case system, derivation patterns, and context-dependent parsing rules compete for working memory simultaneously. Conventional instruction adds extraneous load through extended verbal explanation and mechanical drills, without raising the germane load that builds transferable schemas. PBL repositions students inside a problem they must actively diagnose, increasing germane load and cutting the extraneous load of passive exposure. The Spinning Wheel removes the anxiety cost of the problem's first encounter — lowering affective load before cognitive engagement begins.

2.2 Gamification and Interactive Media in Language Learning

Gamification — game elements applied to non-game instructional settings — has a growing empirical base. Chen et al.'s (2023) meta-analysis across 41 samples (N > 5,000) found a large overall effect (g = 0.82, 95% CI [0.57, 1.08]) on learning outcomes, with motivation and engagement as the most consistent gains. Subhash and Cudney (2018) confirmed that randomisation and immediate feedback reliably reduce learning anxiety in both face-to-face and technology-mediated contexts.

The Spinning Wheel works through unpredictability: students cannot anticipate which problem the wheel will land on, so pre-planned avoidance fails. This keeps attention online and activates curiosity. Psycholinguistically, this maps onto what Krashen (1985, cited in Satria et al.,

2026) termed the affective filter: anxiety blocks acquisition at the input stage; a game situation lowers it, opening the channel. Huda (2020) applied this logic to nahwu specifically and found the wheel effective at making abstract case-marking problems tractable. Ahzan (2025) and Ramadhan (2023) extended the evidence to other Arabic learning contexts.

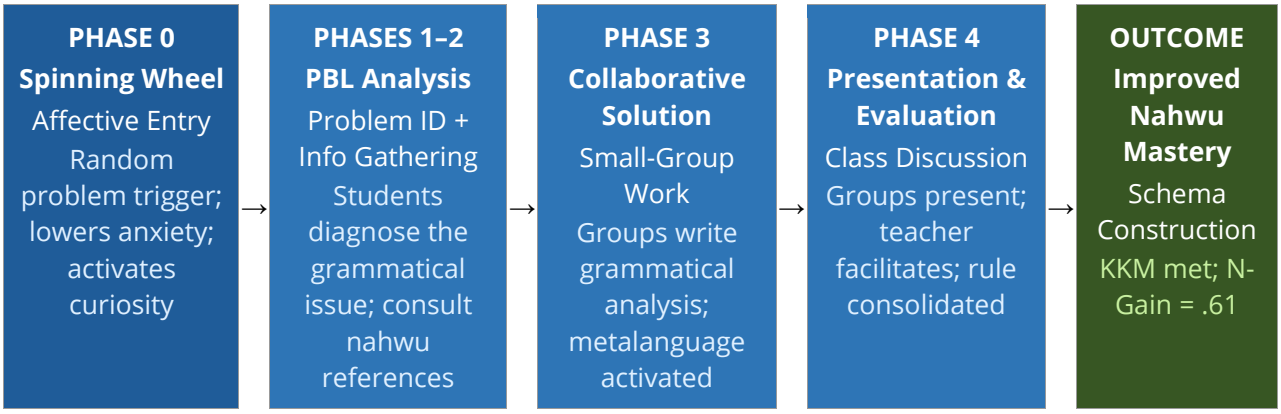
2.3 Nahwu Instruction in Boarding-School Settings

The challenges of nahwu teaching in Indonesian pesantren and madrasah are documented consistently. Wahyudin (2025) and Fatahillah and Wahyudin (2025) identified lecture-based methods and the absence of active-learning structures as the primary drivers of underachievement. Marwaji et al. (2025) showed that visual scaffolds improve grammatical comprehension over text-only explanation. Makinuddin and Amrulloh (2026) found that the Al-Miftah method — which stages grammar-learning through guided problem-solving — raised rule comprehension at madrasah level. The consistent pattern: active problem-engagement structures improve nahwu outcomes, and the mechanism of introduction matters as much as the structure itself.

2.4 Conceptual Framework

Figure 1 maps the study's conceptual model. The Spinning Wheel functions as Phase 0 of PBL: the affective-entry gate. It randomly selects a classical Arabic sentence containing a grammatical problem (case error, agreement mismatch, or iʿrāb ambiguity) and displays it on the classroom screen. This event triggers Phase 1 (Problem Identification), after which students move through information-gathering (Phase 2), collaborative solution-building (Phase 3), and class presentation and evaluation (Phase 4). The expected pathway: wheel randomisation → lowered affective filter → engaged PBL analysis → grammatical schema construction → improved mastery.

Figure 1. *Conceptual Framework: PBL–Spinning Wheel Hybrid Model for Nahwu Instruction*



Theoretical basis: Cognitive Load Theory (Sweller, 1988) • PBL Learning Theory (Hmelo-Silver, 2004) • Affective Filter Hypothesis (Krashen, 1985) • Sociocultural Learning (Vygotsky, 1978)

3. Method

3.1 Research Design and Its Limitations

The study used a quasi-experimental pretest–posttest nonequivalent control-group design (Campbell & Stanley, 1963). Random assignment was not possible because classes are pre-formed within the school's single-sex structure. The most serious validity threat requires explicit treatment here rather than a note in the limitations section. The experimental group (Class IX-1) is entirely female; the control group (Class IX-2) is entirely male. This is a structural confound, not a background demographic detail: gender is nested within group assignment, and ANCOVA — which adjusts for a quantitative pretest covariate — cannot separate the gender effect from the treatment effect because they co-occur perfectly at the class level. To state the interpretive implication plainly: any posttest advantage observed in the experimental group could, in principle, reflect characteristics associated with all-female collaborative groups in Indonesian boarding-school contexts — such as higher rates of peer discussion initiation and metalanguage use (Storch, 2002) — rather than the PBL–Spinning Wheel intervention itself. As a partial transparency measure, the study reports the pretest gap (5.6 points, $p = .091$, nonsignificant) alongside the ANCOVA-adjusted posttest difference (partial $\eta^2 = .585$). The persistence of a large effect after pretest adjustment suggests the group difference is not fully explained by initial ability alone, but this does not rule out gender as an alternative explanation. All findings are therefore framed as preliminary associations, not as evidence of a causal intervention effect, and readers should weight them accordingly.

3.2 Research Site and Participants

The study took place at Al-Hikmah Islamic Boarding School (Maʿhad Al-Ḥikmah), Sumbari, Padang Rajo, West Pasaman, West Sumatra, during the 2025–2026 academic year. Class IX-1 (experimental; $n = 25$, female) and Class IX-2 (control; $n = 25$, male) were selected purposively on the basis of comparable initial nahwu achievement. Both groups had identical weekly Arabic instruction hours, the same institutional curriculum, and three years of prior nahwu study.

Table 1. *Participant Characteristics*

Characteristic	Experimental (IX-1)	Control (IX-2)
n	25	25
Gender	Female (all)	Male (all)
Age (mean years)	14.8	14.9
Years of Arabic study	3	3
Weekly nahwu sessions	3 × 70 min	3 × 70 min
Baseline KKM status	Below (M = 49.4)	Below (M = 55.0)
Gender confound	Yes — major limitation	Yes — major limitation

Source: Authors' compilation based on school records, Al-Hikmah Islamic Boarding School, West Pasaman (2025–2026).

3.3 Data Collection

The study was conducted with the formal permission of Al-Hikmah Islamic Boarding School's principal and academic committee, obtained prior to data collection. Teachers and students were briefed on the study's purpose, procedures, and their right to withdraw at any time without academic consequence. Written informed consent was obtained from teachers; given that all participants were minors, written parental or guardian consent was secured for every student before any data were gathered. No student names appear in any data records or reported outputs: all participants are identified by group label and anonymous code only. The study did not involve clinical procedures, deception, or any risk beyond normal classroom instruction. Because institutional ethics review boards are not yet standard in the Indonesian boarding-school context, the researchers followed the ethical principles of the Indonesian National Research and Innovation Agency (BRIN) and the American Educational Research Association (AERA Council, 2011) guidelines for educational research involving human participants.

3.4 Instrument Development and Validation

The nahwu mastery test comprised 30 items: 10 multiple-choice items on case-marking identification, 10 on agreement analysis, and 10 short-answer items requiring iʿrāb (parsing) of classical Arabic sentences drawn from standard pesantren texts (Matan al-Ājurrūmiyyāh and Al-Imritī). Two expert judges — a lecturer in Arabic linguistics and an experienced madrasah nahwu teacher — independently reviewed all 30 items for content validity, with their judgements interpreted against Lawshe's (1975) CVR framework as a structured reference point (minimum CVR = .62 for two-judge panels); three items fell below the threshold and were revised following their feedback. Given that only two judges participated, the CVR values are treated as indicative of expert agreement rather than definitive psychometric proof of validity. Pilot testing with 20 students outside the sample yielded Cronbach's α = .84. Item difficulty ranged from .38 to .71 (M = .54); discrimination indices ranged from .30 to .62 (M = .45), all meeting the .30 minimum (Ebel & Frisbie, 1991). Because the 10 short-answer iʿrāb items require human judgement in scoring — students must identify the grammatical case, state the governing factor (ʿamil), and supply the correct vowel marking — a three-criteria analytic rubric was developed for each item: (1) correct case identification (1 point), (2) correct ʿamil identification (1 point), and (3) correct vowel marking with justification (1 point), yielding a maximum of 3 points per item (30 points for this section). Two raters independently scored all

pilot-study short-answer responses. Inter-rater reliability was assessed using Cohen's kappa, yielding κ = .82, indicating almost perfect agreement (Landis & Koch, 1977). Discrepancies were resolved through discussion before main-study scoring proceeded.

3.5 Intervention Procedure

Table 2 outlines the eight-session treatment delivered over four weeks. The experimental group followed the PBL–Spinning Wheel protocol (see Figure 1); the control group received conventional direct instruction: teacher explanation, worked example, individual drill.

Table 2. *PBL–Spinning Wheel Session Protocol (Experimental Group)*

Phase	Activity	Duration	Spinning Wheel Role
Phase 0: Problem Trigger	Wheel rotates; a classical Arabic sentence is displayed at random.	10 min	Selects the problem sentence
Phase 1: Problem Identification	Students identify the grammatical issue — case error, agreement mismatch, or iʿrāb ambiguity.	10 min	—
Phase 2: Information Gathering	Groups consult nahwu references to locate the relevant rule.	15 min	—
Phase 3: Collaborative Solution	Groups write a grammatical analysis and correction.	20 min	—
Phase 4: Presentation	Each group presents; the teacher facilitates discussion and provides feedback.	15 min	—

Source: Authors' instructional design adapted from Hmelo-Silver (2004) and Huda (2020).

The Spinning Wheel ran as a browser-based application projected onto the classroom screen. Its item bank held 40 classical Arabic sentences from Matan al-Ājurrūmiyyah and Al-Imritī, each carrying a pre-identified grammatical problem classified by type: case-marking error, ʿadad-maʿdūd agreement, ism-fiʿl agreement, or prepositional-phrase attachment. No sentence was repeated across sessions. Groups of four to five students were kept constant throughout.

3.6 Qualitative Data Collection and Analysis

Qualitative data came from three sources: structured classroom observation (a pre-designed protocol tracking engagement, metalanguage use, and peer interaction per session); semi-structured interviews with the class teacher and six purposively sampled students from the experimental group (three high-achieving, three low-achieving), conducted after the final session; and documentary records including session photographs and student group-work outputs.

Analysis followed Braun and Clarke's (2006) six-phase thematic protocol: familiarisation, initial coding, theme generation, review, definition, and write-up. Two researchers coded

independently; Cohen's kappa was κ = .78, indicating substantial agreement. Three themes emerged and were triangulated with the quantitative results.

3.7 Statistical Analysis

Analysis proceeded in four steps. First, Shapiro–Wilk normality tests were run for all four group-time distributions. Second, Levene's test assessed homogeneity of variance at posttest. Third, an independent-samples t-test compared posttest means. Fourth, ANCOVA was conducted with posttest score as the dependent variable and pretest score as covariate. Before interpreting the ANCOVA, the homogeneity-of-regression-slopes assumption was checked by testing the group × pretest interaction term; a nonsignificant interaction ($p > .05$) confirms that the pretest covariate relates equally to posttest scores in both groups, validating ANCOVA use. Effect size was computed as Cohen's d for the t-test and partial η^2 for ANCOVA. Normalised gain used the formula N-Gain = (post − pre) / (100 − pre). All computations used SPSS 22.

4. Findings and Discussion

4.1 Baseline Equivalence: Pretest Results

Table 3 presents pretest descriptive statistics. Both groups scored below the mastery criterion (KKM = 70). The independent-samples t-test on pretest scores returned $t(48) = 1.73$, $p = .091$, confirming no statistically significant baseline difference despite the 5.6-point gap. Levene's test showed equal variance ($F = 0.41$, $p = .52$). The control group's marginally higher starting mean may partly reflect the uncontrolled gender variable, acknowledged throughout as a residual interpretive constraint.

Table 3. *Pretest Descriptive Statistics*

Indicator	Experimental (IX-1)	Control (IX-2)
n	25	25
Mean (M)	49.4	55.0
SD	9.12	8.74
Minimum	40	40
Maximum	70	75
Pretest t-test	t(48) = 1.73, p = .091	—
Levene's test	F = 0.41, p = .52	—
KKM status	Below (≤ 70)	Below (≤ 70)

Source: Primary data; statistical computation by authors using SPSS 22.

4.2 Normality and Homogeneity Checks

Table 4 reports Shapiro–Wilk results. Three of the four distributions were normal; the experimental group's pretest showed marginal deviation ($W = .94$, $p = .08$). With equal sample sizes ($n = 25$ per group), the central limit theorem supports proceeding with parametric tests. Levene's test at posttest confirmed equal variance ($F = 1.87$, $p = .18$).

Table 4. *Shapiro–Wilk Normality Test Results*

Data	W	p	Status

Pretest — Experimental	.940	.080	Marginal
Pretest — Control	.971	.630	Normal
Posttest — Experimental	.962	.450	Normal
Posttest — Control	.958	.380	Normal

Source: Primary data; SPSS 22 output.

4.3 ANCOVA Assumption: Homogeneity of Regression Slopes

Before interpreting the ANCOVA result, the homogeneity-of-regression-slopes assumption was verified. The interaction between the group (experimental vs. control) and the pretest covariate was added to the ANCOVA model and tested. The interaction term was not significant: $F(1, 46) = 0.43$, $p = .51$. This indicates that the relationship between pretest and posttest scores does not differ across groups — the covariate slopes are parallel — confirming that ANCOVA is an appropriate analytical strategy for this data. The pretest was then removed from the interaction term and ANCOVA was re-run with pretest as a simple covariate.

4.4 Posttest Results

After the eight-session intervention, the experimental class outperformed the control by a wide margin. Table 5 presents the posttest descriptive statistics, and Figure 2 displays these as a grouped bar chart with error bars representing one standard deviation.

Table 5. *Posttest Descriptive Statistics*

Indicator	Experimental (IX-1)	Control (IX-2)
n	25	25
Mean (M)	80.8	62.6
SD	7.46	10.42
Minimum	65	45
Maximum	95	85
KKM met (≥70)?	Yes (group mean)	No (group mean)
Performance category	Good–Excellent	Adequate

Source: Primary data; statistical computation by authors using SPSS 22.

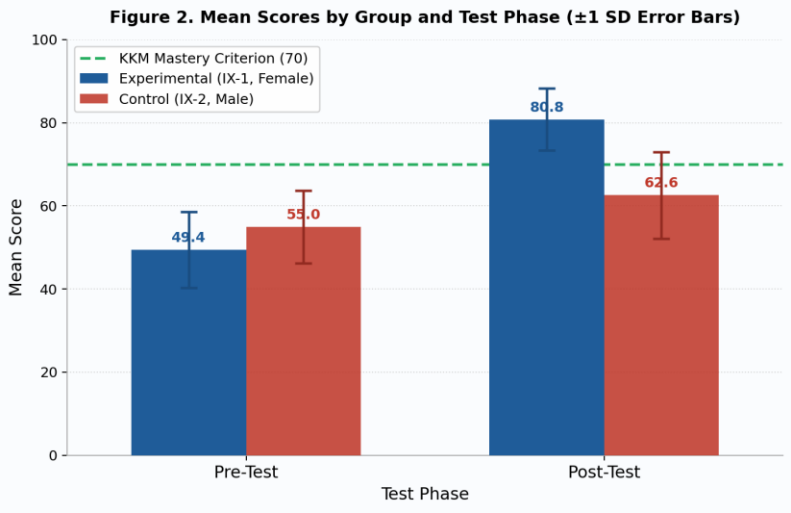


Figure 2. Group Mean Scores at Pretest and Posttest with ±1 SD Error Bars. The experimental group crossed the KKM threshold (70) at posttest; the control group did not.

The experimental group's mean rose 31.4 points from pre- to posttest; the control group's rose 7.6 points. The smaller SD in the experimental group (7.46 vs. 10.42) suggests more even performance, though ceiling effects for the top scorers cannot be excluded.

4.5 Inferential Statistics and Effect Sizes

Table 6 presents the full inferential results. Both the t-test and ANCOVA indicate a large and statistically significant group difference. Given the uncontrolled gender confound, these statistics are evidence of a strong association, not proof of causation.

Table 6. *Posttest Inferential Statistics*

Statistic	Value	Note
t-test	$t(48) = 8.14$, $p < .001$	Large group difference
Mean difference	18.2 points	Exp. − Control
95% CI	[13.6, 22.8]	Excludes zero
Cohen's d	2.01	Large (Cohen, 1988)
Levene's (posttest)	$F = 1.87$, $p = .18$	Equal variances confirmed
Regression slopes test	$F(1, 46) = 0.43$, $p = .51$	HRS assumption met
ANCOVA F	$F(1, 47) = 66.3$, $p < .001$	Pretest controlled
Partial η^2	0.585	Large effect size
ANCOVA Adjusted Means	Exp. = 79.1; Control = 64.3	Pretest covariate adjusted; difference narrows slightly vs. unadjusted
Gender confound	Not controlled in ANCOVA	Major validity constraint

Source: Primary data; independent-samples t-test and ANCOVA computed via SPSS 22.

4.6 Normalised Gain (N-Gain)

Table 7. *Normalised Gain by Group*

Group	Pre M	Post M	N-Gain	Category (Hake, 1998)
Experimental (IX-1)	49.4	80.8	.61	Medium–High
Control (IX-2)	55.0	62.6	.17	Low

Source: Primary data; N-Gain computed using the formula (post − pre) / (100 − pre); categories follow Hake (1998).

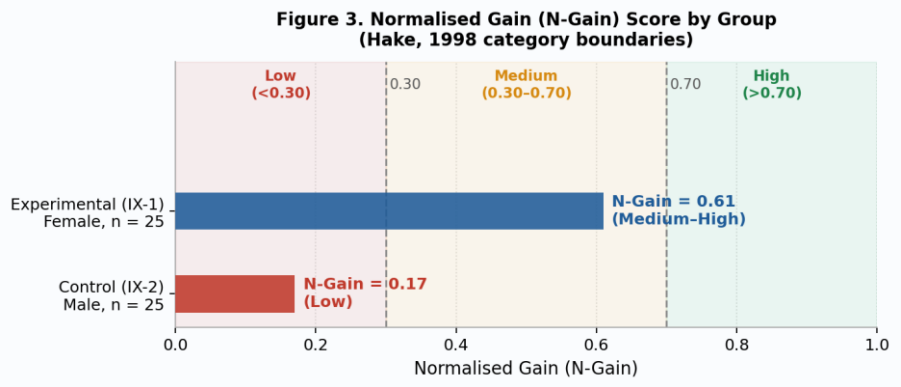


Figure 3. N-Gain Score by Group. Experimental group (N-Gain = .61) reaches medium–high range; control group (N-Gain = .17) remains low. Boundaries follow Hake (1998).

The experimental group's N-Gain of .61 captures 61% of the maximum possible improvement above its pretest baseline. The control group's .17 falls in the low category, reflecting that conventional instruction produced limited growth relative to ceiling.

4.7 Qualitative Findings

Thematic analysis produced three themes (Cohen's κ = .78).

Theme 1: Affective engagement through randomised challenge. Spontaneous discussion began within 30 seconds of problem display in all eight sessions. The class teacher described the shift: "After the first session with the wheel, the students stopped asking 'what are we supposed to do?' — they started asking 'what does this sentence mean?'" Student interviews corroborated this. One student in the low-achieving cohort reported: "When the wheel spins, I feel like I want to answer first. I forget that I might be wrong." A high-achieving student added: "Before this, I always waited for the teacher to explain. Now I try to figure it out with my group." These self-reports are not validated anxiety measures, but they are consistent with the affective filter mechanism and with the observed engagement data.

Theme 2: Metalanguage activation in peer discussion. Observation coding showed that experimental-group students used grammatical metalanguage — marfu', manṣūb, majrūr, fāʿil, mafʿūl bih — far more frequently during group work than control-group students did in paired practice. Written group outputs followed a structured analytical chain — problem identification → rule citation → correction — that was largely absent from control-group written work.

Theme 3: Self-directed error noticing. Four experimental-group students described noticing and correcting grammatical errors in independent reading outside class. One student said: "I read a sentence in my kitab last night and I noticed the iʿrāb was wrong — I could see it because we had analysed something similar in class." This is consistent with transfer but cannot be formally measured in this design. The theme is treated as interpretive and requires a delayed posttest to validate.

Qualitative data were collected from the experimental group only. The themes corroborate the quantitative results and offer process-level illustration of the probable mechanism, but they do not constitute independent triangulation.

4.8 Interpreting the Results

The results speak to both research questions. RQ1 is answered by the intervention protocol in Section 3.4 and Figure 1: the model ran through eight structured sessions in which wheel rotation initiated each lesson and PBL phases provided the analytical architecture. RQ2 is answered by the data above: the model is associated with substantially improved nahwu mastery, with converging evidence from the t-test, ANCOVA, N-Gain scores, and qualitative themes.

Before drawing interpretive conclusions, the validity constraint must be stated without softening. Cohen's d of 2.01 is large by any benchmark (Cohen, 1988), but it cannot be attributed to the intervention alone. Research on gender and collaborative language learning has documented that all-female groups often show higher rates of peer interaction and discussion initiation than all-male groups (Storch, 2002). This study cannot disentangle the intervention effect from the gender effect. The reported effect size should be read as an upper bound, not a precise estimate of what the model would yield under gender-balanced conditions.

With that constraint in view, the theoretical account of the observed association is still analytically useful. Cognitive load theory (Sweller, 1988) predicts that designs reducing extraneous load should improve outcomes in high-intrinsic-load domains — classical Arabic grammar is precisely that. The Spinning Wheel turns the most cognitively threatening moment in a nahwu lesson — first contact with a grammatical problem — into a game-initiated challenge. Chen et al.'s (2023) meta-analysis establishes that this kind of gamified entry reliably shifts engagement in ways that translate to learning outcomes. PBL then channels that arousal into germane cognitive work: students argue about case assignments, deploy metalanguage, and construct grammatical schemas through active problem-solving. Hmelo-Silver (2004) showed that this process produces more flexible and transferable knowledge than direct instruction.

Three rival explanations deserve acknowledgment. The novelty effect is real: the Spinning Wheel was new at this school, and some engagement advantage almost certainly reflects initial curiosity. A delayed posttest is needed to separate this from durable learning. The teacher effect is live: the same teacher delivered both conditions, but in structurally different modes; differential enthusiasm for PBL cannot be excluded. The gender effect, already foregrounded above, is the most serious: future designs must balance gender across conditions.

What the study suggests, within these limits, is that classical nahwu content and contemporary active pedagogy are not in competition. The model used texts already in the curriculum — Matan al-Ājurrūmiyyah, Al-Imritī — and changed students' relationship to that content, from reception to active problem-solving. Akzam and Hayati (2026) called this "balanced innovation"; the present study offers one of the first quasi-experimental data points bearing on that claim in a boarding-school setting.

5. Conclusion

This study found that combining PBL with Spinning Wheel media was associated with substantially improved nahwu mastery among ninth-grade students at Al-Hikmah Islamic Boarding School. The experimental group's posttest mean (80.8) crossed the mastery threshold and exceeded the control group's mean (62.6): $t(48) = 8.14$, $p < .001$, Cohen's $d = 2.01$, N-Gain = .61. ANCOVA — validated by the homogeneity-of-regression-slopes check — confirmed the group difference after accounting for the pretest gap: $F(1, 47) = 66.3$, $p < .001$, partial $\eta^2 = .585$. Qualitative evidence added process-level texture: higher affective engagement, more frequent metalanguage use in peer discussion, and student-reported error noticing. These results are treated as preliminary evidence of promise, not as confirmation of effectiveness, given the quasi-experimental design and the unresolved gender confound.

For teachers and curriculum designers in Islamic boarding schools, the study points to one actionable possibility: classical nahwu content and active pedagogy can coexist in the same lesson. The PBL–Spinning Wheel model slots into existing structures without touching the syllabus. The wheel draws from the same classical texts already in use; PBL replaces teacher-fronted explanation with student-centred problem-solving over the same grammatical material. Based on the implementation, the approach appears to work best when wheel problems have a clearly identifiable grammatical issue, groups are four to five students, and the teacher holds back corrections until groups have made their own attempt. Eight sessions over four weeks produced a measurable shift; whether a longer run sustains or amplifies it remains open.

Four limitations shape what can be concluded. First, and most seriously, the all-female experimental group and all-male control group represent an uncontrolled confound that prevents causal attribution. Some share of the outcome gap may reflect gender-linked differences in collaborative learning rather than the intervention itself. Second, the sample is one school, one grade, one semester. Third, affective variables were not measured with validated instruments; the affective filter account remains interpretive. Fourth, the absence of a delayed posttest leaves the durability of gains unknown.

Future research should use gender-balanced random assignment or a matched split-class format; replicate across school types, grade levels, and Arabic text traditions; incorporate validated affective measures such as the Foreign Language Classroom Anxiety Scale adapted for Arabic (Horwitz et al., 1986); include a delayed posttest at six to eight weeks; and collect parallel qualitative data from both groups to make genuine triangulation possible.

References

Adiana, W. N. (2024). Pengaruh Model Pembelajaran Problem-Based Learning Berbantuan Media Spinning Wheel Terhadap Berpikir Kritis Siswa Pada Mata Pelajaran Sosiologi [PhD Thesis]. Universitas Mataram.

AERA Council. (2011). Code of Ethics American Educational Research Association Approved by the AERA Council. https://www.aera.net/about-aera/aera-rules-policies/professional-ethics

Ahzan, H. A. (2025). Efektivitas Penggunaan media "Spinning Wheel" dalam pembelajaran 'Adad dan Ma'dud di Dayah Modern Al-Furqan Bireuen. Fathir: Jurnal Studi Islam, 2(3), 445–458. https://doi.org/10.71153/fathir.v2i3.377

Akzam, I., & Hayati, N. (2026). Integrasi Pendidikan Bahasa Arab Tradisional dan Metode Modern dalam Kurikulum Bahasa Arab Murni. Jurnal Dedikasi Pengabdian Pendidikan, 2(1), 15–29. https://doi.org/10.64008/jdpp.v2i1.53

Al-Drees, A. A., Al-Rubaish, A. M., & Al-Muhanna, F. A. (2024). Effects of problem-based learning on EFL learning: A systematic review. PLOS ONE, 19(12), e0307819. https://doi.org/10.1371/journal.pone.0307819

Barrows, H. S., & Tamblyn, R. M. (1980). Problem-based learning: An approach to medical education. Springer.

Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. Qualitative Research in Psychology, 3(2), 77–101. https://doi.org/10.1191/1478088706qp063oa

Campbell, D. T., & Stanley, J. C. (1963). Experimental and quasi-experimental designs for research. Houghton Mifflin.

Chen, F., Lui, A. M., & Martinelli, S. M. (2023). Examining the effectiveness of gamification as a tool promoting teaching and learning in educational settings: A meta-analysis. Frontiers in Psychology, 14, 1253549. https://doi.org/10.3389/fpsyg.2023.1253549

Cohen, J. (1988). Statistical power analysis for the behavioral sciences (2nd ed.). Lawrence Erlbaum Associates.

Cook, V. (2016). Second language learning and language teaching (5th ed.). Routledge.

Dolmans, D. H. J. M., Loyens, S. M. M., Marcq, H., & Gijbels, D. (2015). Deep and surface learning in problem-based learning: A review of the literature. Advances in Health Sciences Education, 21(5), 1087–1112. https://doi.org/10.1007/s10459-015-9645-6

Ebel, R. L., & Frisbie, D. A. (1991). Essentials of educational measurement (5th ed.). Prentice-Hall.

Ellis, R. (2006). Current issues in the teaching of grammar: An SLA perspective. TESOL Quarterly, 40(1), 83–107. https://doi.org/10.2307/40264512

Fatahillah, A., & Wahyudin, D. (2025). Analisis penyebab kesulitan belajar nahwu pada santri madrasah diniyah Banul Ishlah Al-Ibrahimiyah Perampuan Barat Kecamatan Labuapi. J-Simbol: Jurnal Magister Pendidikan Bahasa Dan Sastra Indonesia, 13(1), 380–386. https://doi.org/10.23960/simbol.v13i1.560

Fauzi, M. S. (2020). Pengaruh model pembelajaran Problem-Based Learning terhadap hasil belajar nahwu siswa kelas X SMA Sains Wahid Hasyim Sleman 2019/2020 [Unpublished undergraduate thesis]. Universitas Islam Negeri Sunan Kalijaga.

Hmelo-Silver, C. E. (2004). Problem-based learning: What and how do students learn? Educational Psychology Review, 16(3), 235–266. https://doi.org/10.1023/B:EDPR.0000034022.16470.f3

Horwitz, E. K., Horwitz, M. B., & Cope, J. (1986). Foreign language classroom anxiety. The Modern Language Journal, 70(2), 125–132. https://doi.org/10.1111/j.1540-4781.1986.tb05256.x

Huda, N. F. (2020). Penggunaan media pembelajaran Spinning Wheel dalam pembelajaran qawāʿid nahwu. Jurnal Pendidikan Bahasa Arab, 4(2), 45–58.

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. Biometrics, 33(1), 159–174. https://doi.org/10.2307/2529310

Lawshe, C. H. (1975). A quantitative approach to content validity. Personnel Psychology, 28(4), 563–575. https://doi.org/10.1111/j.1744-6570.1975.tb01393.x

Makinuddin, M., & Amrulloh, F. S. (2026). Penerapan metode Al-Miftah dalam meningkatkan pemahaman kaidah nahwu di Madrasah Tsanawiyah Mambaus Sholihin Gresik. TADRIS AL-ARABIYAT: Jurnal Kajian Ilmu Pendidikan Bahasa Arab, 6(1), 93–110.

Marwaji, D., Koderi, K., Akmansyah, M., Amiruddin, A., & Erlina, E. (2025). A graph-based grammar learning model for Salafi Islamic boarding school media: A literature review in SINTA-accredited journals during the period 2018–2024. Alsuna: Journal of Arabic and English Language, 8(1), 235–248. https://doi.org/10.31538/alsuna.v8i1.7485

Ramadhan, R. M. (2023). Peningkatan prestasi fahm al-maqrū dan kafā'ah al-kitābah dalam pembelajaran bahasa Arab dengan menggunakan Spinning Wheel pada peserta didik di Rumah Yatim Yogyakarta Cabang Monjali [Doctoral dissertation]. Universitas Islam Indonesia.

Savery, J. R. (2006). Overview of problem-based learning: Definitions and distinctions. Interdisciplinary Journal of Problem-Based Learning, 1(1), 9–20. https://doi.org/10.7771/1541-5015.1002

Seff, F. M. (2019). Dinamika pendidikan bahasa Arab di Indonesia dalam konteks persaingan global. IAIN Antasari Press.

Storch, N. (2002). Patterns of interaction in ESL pair work. Language Learning, 52(1), 119–158. https://doi.org/10.1111/1467-9922.00179

Subhash, S., & Cudney, E. A. (2018). Gamified learning in higher education: A systematic review of the literature. Computers in Human Behavior, 87, 192–206. https://doi.org/10.1016/j.chb.2018.05.028

Subhi, I. M., Putri, R. W., Mahmudah, S., Anshory, I., & Restiani, A. (2026). Problem-Based Learning as an instructional strategy for strengthening students' literacy and numeracy skills: A systematic literature review. Santhet: Jurnal Sejarah Pendidikan Dan Humaniora, 10(2), 88–104.

Sweller, J. (1988). Cognitive load during problem solving: Effects on learning. Cognitive Science, 12(2), 257–285. https://doi.org/10.1016/0364-0213(88)90023-7

Wahyudin, D. (2025). Analisis kesulitan belajar nahwu di lembaga pendidikan pesantren: Studi kasus di Jawa Barat. J-Simbol: Jurnal Magister Pendidikan Bahasa Dan Sastra Indonesia, 13(2), 201–215.