The Limits of Problem-Based Learning in *Arabic* Grammar: *Na'at–Man'ūt* Achievement and Participation Inequality in a Quasi-Experiment

Moh Sholeh Afyuddin[1]*, Muflihah Rahayu[2], M. Syamsul Ma'arif[3], Munjiyat[4]
[1,2,3]Program Studi Pendidikan Bahasa Arab, Universitas Islam Negeri Syekh Wasil Kediri, Indonesia,
[4]Faculty of Letters and Human Sciences, Mohammed V University in Rabat, Morocco
sholehafyuddin@gmail.com, muflikhahrahayu57@gmail.com, msymaarif@gmail.com, munjiyat@gmail.com
**Corresponding Author*

DOI: https://doi.org/10.62730/syaikhuna.v17i2.7897
Article History: Received: [14-08-2026] Revised: [08-09-2026] Accepted: [08-09-2026] Published: [29-10-2026]
Copyright & Licensing: Copyright © [2026] Moh Sholeh Afyuddin et al. This article is licensed under a Creative Commons Attribution 4.0 International License (CC BY 4.0). https://creativecommons.org/licenses/by/4.0/

Abstract

This study examines the friction between unscaffolded Problem-Based Learning (PBL) and Arabic grammar (qawaid) instruction, specifically the na'at–man'ūt construction, among eleventh-grade students. A quasi-experimental pretest–posttest control group design was employed with 58 students in two intact classes selected via purposive sampling. Quantitative data were collected through a 20-item grammar test and analyzed using the Mann–Whitney U test; qualitative data were gathered through a structured observation protocol coded by two independent raters (κ = .78). While the PBL class descriptively outperformed the lecture-based class in N-Gain (0.6413 versus 0.4755), inferential testing yielded a non-significant result (U = 381.50, p = .479) with a small-to-moderate effect size (d = 0.47, 95% CI [−0.05, 0.99]). It achieved only .38 power; therefore, the study is inconclusive regarding differences in achievement. Observations revealed a widening participation gap: high achievers increased their verbal turns from 57.2% to 68.4%. In comparison, low achievers declined from 10.3% to 8.4%, consistent with differential cognitive load but not direct proof of overload. A four-phase Scaffolded-PBL model is proposed as a testable hypothesis for future research, not as a validated intervention.

Keywords: *Problem-Based Learning; Arabic Grammar; Element Interactivity; Participation Inequality; Quasi-Experimental Design*

INTRODUCTION

The cognitive architecture of Arabic grammar (*qawaid*) instruction presents a persistent pedagogical challenge that extends well beyond mere methodological preference. For novice learners, mastering the *na'at* and *man'ūt* system—governing adjectival agreement across gender (*jins*), number (*'adad*), definiteness (*ta'rīf/tankīr*), and case inflection (*i'rāb*)—imposes an exceptionally high element interactivity demand on working memory. According to Cognitive Load Theory (CLT), intrinsic cognitive load is determined by the number of interacting information elements that must be processed simultaneously; when these elements cannot be meaningfully chunked into established cognitive schemas, working memory capacity is rapidly exhausted (Sweller, 2020; Sweller et al., 2023). As a core construct of CLT, element interactivity is the primary mechanism that generates and modulates intrinsic and extraneous loads during complex skill acquisition (de Jong, 2010; Leahy & Sweller, 2019). The *na'at–man'ūt* construction requires learners to coordinate four distinct grammatical parameters in real time: noun

definiteness (*ma'rifah* vs. *nakirah*), morphological gender (*mudhakkar* vs. *mu'annath*), numerical state (*mufrad*, *muthannā*, *jam'*), and syntactic case (*raf'*, *naṣb*, *jarr*). A single error in any parameter triggers a cascading violation across the entire phrase. This is not merely a matter of rule complexity; it imposes a heavy simultaneous processing burden that novice learners cannot chunk without explicit schema activation (Skulmowski, 2021; Krieglstein et al., 2025). Furthermore, recent multidimensional cognitive load research underscores that when novices face high intrinsic complexity without initial structural guidance, extraneous cognitive load surges, further impairing germane schema formation (Antonenko et al., 2010; Andersen & Makransky, 2021).

The pedagogical consequences of this rigid cognitive architecture are well documented across language education. Learners in Indonesian madrasah and *pesantren* institutions consistently report *qawaid* as the most intimidating and anxiety-inducing component of Arabic learning (Fauzi, 2021; Abu Yousef & Abu Owda, 2026), a perception perpetuated by traditional instructional traditions that prioritize rote rule memorization over functional syntactic application (Abdurrohman & Ulinnuhaa, 2024). Students frequently recite abstract morphological rules with high verbal accuracy, yet become cognitively paralyzed when tasked with constructing or parsing authentic sentences in real time. This persistent dissociation between declarative rule recall and procedural syntactic application is precisely what skill acquisition theory predicts when instruction neglects the systematic transition from explicit rule awareness to automatized execution (Afiyah, 2025; Boulaares, 2026). Recent systematic evidence on Arabic language processing confirms that without structured instructional bridging, cognitive friction, reading hesitation, and syntactic processing disabilities remain prevalent among non-native learners (Mohammed et al., 2025). While recent scholarship emphasizes the necessity of integrated grammatical instruction, the precise cognitive mechanisms governing how novice learners process multi-variable grammatical agreement under active learning conditions remain insufficiently explored.

To overcome the limitations of traditional grammar instruction, Problem-Based Learning (PBL) has been widely championed as a transformative pedagogical paradigm grounded in social constructivism. Because knowledge is best constructed through authentic, collaborative problem-solving, PBL engages learners in diagnosing ill-structured tasks and fosters critical thinking, metacognitive reflection, and self-directed learning skills (Hmelo-Silver, 2004; Choo, 2014). In second language acquisition (SLA), PBL has been increasingly advocated to shift instruction from passive drill-and-practice toward authentic communicative negotiation and contextual rule discovery (Allen et al., 2011; Al-Busaidi et al., 2021). However, directly transplanting unscaffolded PBL into high-element-interactivity grammatical domains creates serious theoretical friction. Empirical investigations into language teacher practice reveal widespread resistance to unguided PBL, driven by teachers' recognition that open-ended inquiry frequently induces cognitive disorientation in grammar classrooms (Moslemi Nezhad Arani et al., 2023). When novices are required to deduce underlying syntactic patterns without prior schema support independently, unguided discovery demands extensive random search through working

memory. This unassisted search through the problem space generates immense extraneous cognitive load, exhausting limited mental capacity before productive schema acquisition can occur (Kirschner et al., 2006; Sweller, 2020).

To resolve this instructional dilemma, cognitive scientists and educational researchers emphasize the crucial mediating role of instructional scaffolding and systematic task sequencing. Scaffolding—temporary structural support provided by expert facilitators or structured tools—enables learners to perform complex tasks beyond their unassisted capacity (Belland et al., 2017). Meta-analytic synthesis confirms that structured scaffolding significantly enhances cognitive outcomes across problem-centered learning environments (Belland et al., 2017; Ak, 2016). Specifically, research on cognitive load management shows that sequencing content into manageable sub-tasks and gradually fading instructional support minimizes extraneous processing, freeing working memory for germane schema construction (Lange & Costley, 2022; Gidalevich & Kramarski, 2018). In the SLA literature, Spada and Tomita's (2010) meta-analysis shows that explicit, scaffolded instruction is particularly vital for complex grammatical features, whereas unguided exposure yields negligible benefits for novice learners. Empirical evidence further demonstrates that integrating explicit grammar modeling with scaffolded practice bridges the gap between declarative rule knowledge and procedural fluency in L2 production (Moeen et al., 2019; Ahmadi Safa & Motaghi, 2021). Without deliberate scaffolding, unguided inquiry fails to provide the cognitive architecture novices need to navigate complex morphosyntactic structures.

The cognitive friction generated by unscaffolded PBL extends beyond individual working memory constraints to influence social interaction dynamics and classroom participation equity. When novice learners are confronted with high-interactivity tasks that exceed their working memory capacity, heightened cognitive load frequently triggers affective distress and foreign language anxiety (Abu Yousef & Abu Owda, 2026; Zhang & Saisai, 2026). In collaborative group settings, unguided problem-solving can exacerbate social inequality and participation disparities (Liu, 2026). When group inquiry lacks explicit role structuring or adaptive scaffolding, high-achieving students with stronger baseline schemas dominate verbal discussions, while lower-achieving peers experience cognitive overload, performance anxiety, and eventual task withdrawal (Qutieshat et al., 2020; Rababah et al., 2025). This trajectory converts active learning groups into exclusive domains for high achievers, fostering system dependency and passive withdrawal among struggling learners (Frambach et al., 2014; Rababah et al., 2025). Thus, evaluating active learning models requires examining not only post-instruction achievement gains but also the underlying real-time participation dynamics that reflect learners' cognitive and affective states during problem-solving.

This tension—between the constructivist ideal of open problem-solving and novice learners' cognitive architecture—is what this study terms the *Procedural-Constructivist Conflict*. Recent experimental studies show that for low-prior-knowledge learners, explicit scaffolding before problem-solving yields better achievement than inquiry-first approaches, illustrating the expertise reversal effect (Zambrano et al., 2026; Ngu et al., 2026). Scaffolding interventions within

PBL—such as concept maps, worked examples, and structured completion tasks—have been proven to reduce element interactivity and optimize cognitive load (Hendrayana & Mutaqin, 2025; Bozorgian et al., 2026). Despite these theoretical advances, the empirical literature on Arabic grammar instruction lacks systematic mixed-methods studies that evaluate both quantitative achievement outcomes and qualitative participation dynamics in unscaffolded PBL. The novelty of this study lies in its three interrelated contributions: (1) empirically testing the *Procedural-Constructivist Conflict* in Arabic *qawaid* instruction (*na'at–man'ūt* agreement); (2) capturing the participation inequality trajectory across higher and lower achievement groups using structured video observations; and (3) proposing a theoretically calibrated, four-phase *Scaffolded-PBL* model as an actionable framework for madrasah pedagogy. Specifically, this study addresses two research questions: (1) Do post-instruction *qawaid* achievement scores differ significantly between unscaffolded PBL and lecture-based intact classes after accounting for baseline performance? (2) How are verbal participation turns distributed across higher and lower achievement groups during PBL sessions, and what does this reveal regarding cognitive load during collaborative inquiry?

METHOD

Research Design

This study used a quasi-experimental pretest–posttest control-group design with two intact classes, supplemented by structured qualitative observation. The design was selected because random assignment of individual students was not feasible in the existing school structure. At the same time, the pretest–posttest format permits examination of achievement change relative to baseline. The quantitative strand addressed the first research question concerning post-instruction grammar outcomes; the qualitative observation strand addressed the second research question concerning the distribution of verbal participation during PBL sessions. This combination follows a convergent mixed-methods logic in which quantitative and qualitative data are collected in parallel and integrated at the interpretation stage to capture both cognitive outcomes and classroom process dynamics (Creswell & Plano Clark, 2018). The study does not test the four-phase Scaffolded-PBL model; that model is treated as a post-hoc conceptual proposition derived from the findings.

Participants and Sampling

The study was conducted at MAN 1 Kediri during the 2025/2026 academic year. The accessible population comprised eleventh-grade students enrolled in the Islamic Senior High School curriculum. Two intact classes were selected through purposive sampling based on equivalent baseline academic achievement, as confirmed by prior semester Arabic language grades and pretest scores. We assigned Class XI-A (n = 29) to the experimental group, which received PBL instruction, while Class XI-B (n = 29) served as the control group and received lecture-based instruction. The total analytic sample was N = 58. All participants were native

Indonesian speakers learning Arabic as a foreign language and had comparable prior exposure to basic Arabic grammar.

Three limitations of this sampling procedure warrant explicit acknowledgment. First, using intact classes introduces potential confounding between class membership and treatment condition; differences in classroom climate, peer composition, or scheduling cannot be fully separated from the instructional method itself. Second, purposive assignment rather than randomization limits causal inference. Third, the sample size of 58 yields achieved statistical power of approximately .38 for the observed effect, which is insufficient for definitive hypothesis testing and is addressed further in the Data Analysis subsection.

Instruments and Validation

Arabic Grammar Achievement Test (AGAT). The AGAT consisted of 20 dichotomously scored items assessing four sub-dimensions of *naʿat–manʿūt* agreement: definiteness matching (5 items), case ending alignment (5 items), gender agreement (5 items), and number consistency (5 items). Content validity was established by three experts in Arabic language pedagogy and educational assessment, who reviewed item relevance and alignment with the intended construct. Item analysis yielded a Cronbach's alpha reliability coefficient of $\alpha = .84$ for the overall test. For a dichotomously scored instrument, KR-20 would provide a more technically appropriate internal consistency estimate; however, alpha is reported here as the coefficient actually computed, and this choice is acknowledged as a limitation.

Structured Group Observation Protocol (SGOP). The SGOP recorded verbal participation turns, group engagement indicators, and help-seeking behaviors during PBL sessions. We operationally defined a "verbal participation turn" as any audible, task-relevant utterance a student initiated during small-group discussion, bounded by the next speaker's turn or a pause exceeding five seconds. We classified students into two achievement groups (higher and lower) using a quartile split based on their AGAT pretest scores. Students scoring in the top quartile were assigned to the higher-achievement group, while those scoring in the bottom quartile were assigned to the lower-achievement group. This classification was used to examine differences in classroom participation across achievement levels. To establish coding reliability, two independent raters scored a 20% random sample of observation sessions, yielding substantial inter-rater agreement (Cohen's $\kappa = .78$).

Cognitive Load Scale. Although the original research plan included the Leppink et al. (2013) Cognitive Load Scale, scheduling constraints prevented consistent administration across all sessions. Consequently, this article analyzes and reports no cognitive load scores. Any reference to cognitive overload in the Discussion is a theoretical interpretation of observed participation patterns and score dispersion, not an empirically measured finding. This decision aligns with the review feedback that overload claims require validly reported cognitive-load data.

Data Collection Procedures

Data collection proceeded in four stages. First, administrative permissions were obtained from the school principal and classroom teachers, and informed consent was secured from

student guardians. Second, researchers administered the AGAT pretest to both classes in a single 45-minute session during regular class hours. Third, the intervention was implemented over two weeks in three sessions. The experimental group received PBL instruction on the *na'at–man'ūt* construction, structured around ill-structured grammar problems requiring collaborative diagnosis of agreement errors in authentic Arabic sentences. The control group received lecture-based instruction on the same topics, delivered by the same teacher, with explicit rule explanation and worked examples. Both groups used identical learning materials and were allocated equivalent instructional time. Fourth, the AGAT posttest was administered to both classes under the same conditions as the pretest. We conducted structured observations during all three PBL sessions in the experimental class only, because the research question about participation distribution was specific to the PBL environment.

Data Analysis

Quantitative data were analyzed using SPSS 26.0 and R 4.3.1. We calculated normalized gain (N-Gain) scores for each student to index improvement relative to baseline capacity using the formula N-Gain = (posttest − pretest) / (maximum score − pretest). Shapiro–Wilk tests revealed significant departures from normality in the experimental group's posttest and N-Gain scores ($p < .05$), and Levene's test indicated heteroscedasticity across groups ($F = 5.814$, $p = .019$). These violations justified using the non-parametric Mann–Whitney U test for between-group comparisons. Effect sizes were estimated using Cohen's d with 95% confidence intervals. A post-hoc power analysis was conducted using G*Power 3.1.9.7 (Faul et al., 2007) to evaluate the achieved power given the final sample size, alpha level ($\alpha = .05$), and observed effect size. Given the limited power, the non-significant inferential result is interpreted as inconclusive rather than as evidence of no effect.

We analyzed qualitative observation data by descriptively quantifying participation turns by higher- and lower-achievement groups across the three sessions. Turn counts were converted to percentages of total verbal turns per session, and a dominance index (high-achiever turns ÷ low-achiever turns) was computed to characterize participation polarization. This approach does not constitute a full qualitative analysis; it provides a structured process description intended to complement the achievement data.

Ethical Considerations

Institutional review approval was obtained from the ethics committee of IAIN Kediri before data collection (Protocol No. 2025/ETH/PBA-042). Formal consent was secured from school administrators, classroom teachers, and student guardians. Participation was voluntary, and participants were assured of confidentiality and anonymity. Control group students received the same PBL learning materials after completing the experimental intervention to ensure pedagogical equity.

RESULTS AND DISCUSSION

Results

Descriptive Statistics

Table 1 presents the descriptive statistics for pretest, posttest, and normalized gain (N-Gain) scores across the experimental and control groups. Both classes entered the study with comparable baseline performance, as reflected in pretest means of 42.15 (SD = 11.48) for the experimental group and 44.83 (SD = 10.92) for the control group, a difference of less than three points. After the intervention, the experimental group had a posttest mean of 79.28 (SD = 20.26), while the control group had 71.03 (SD = 13.92). The mean posttest difference of 8.25 points descriptively favored the PBL class. Normalized gain followed the same descriptive pattern, with the experimental group at 0.6413 (SD = 0.312) and the control group at 0.4755 (SD = 0.218).

Table 1

Descriptive Statistics of Pretest, Posttest, and N-Gain Scores by Group

Group	n	Pretest M	Pretest SD	Posttest M	Posttest SD	N-Gain M	N-Gain SD
Experimental (PBL)	29	42.15	11.48	79.28	20.26	0.6413	0.312
Control (Lecture)	29	44.83	10.92	71.03	13.92	0.4755	0.218

Note. Scores are expressed on a 0–100 scale. N-Gain was calculated as (posttest − pretest) / (100 − pretest) for each student.

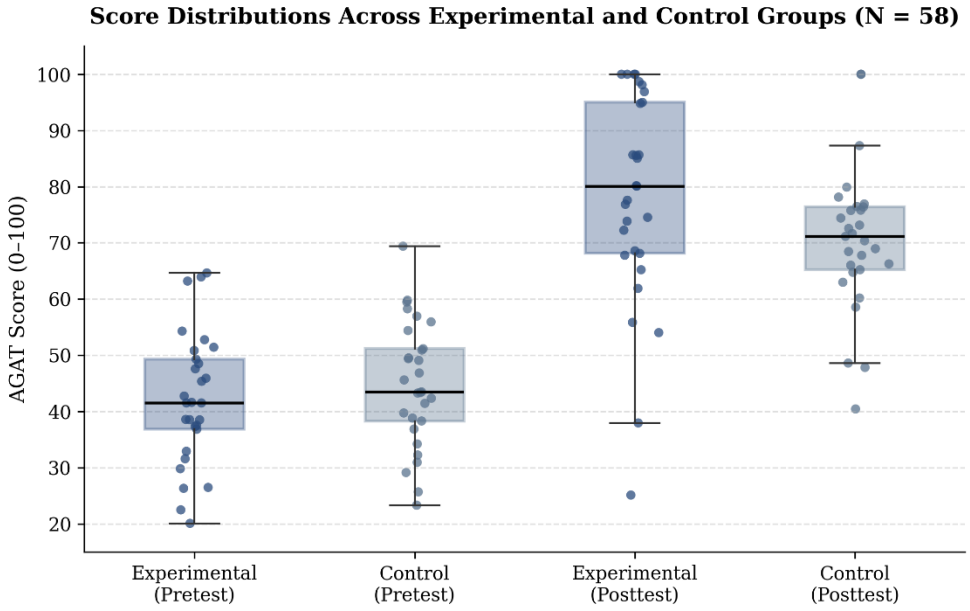


Figure 1

Score Distributions Across Experimental and Control Groups

Note. Side-by-side box plots show the median, interquartile range, and whiskers, with overlaid individual data points (jitter), for the experimental pretest, control pretest, experimental posttest, and control posttest (N = 58).

Prerequisite Tests and Hypothesis Testing

Before inferential testing, assumptions were checked. Shapiro–Wilk tests indicated that experimental group posttest scores departed significantly from normality (W = 0.892, p = .008), while Levene's test for equality of variances indicated heteroscedasticity across groups (F = 5.814,

p = .019). These violations justified using the non-parametric Mann–Whitney U test for the primary between-group comparison rather than a parametric alternative such as ANCOVA.

Table 2 reports the full inferential output. The Mann–Whitney U test yielded U = 381.50, Z = −0.708, p = .479 (two-tailed), with mean ranks of 30.84 for the experimental group and 28.16 for the control group. The effect size was Cohen's d = 0.47, with a 95% confidence interval of −0.05 to 0.99. This interval includes zero, which is consistent with the non-significant test result: the data are compatible with both a small negative effect and a moderate positive effect. The achieved statistical power for this sample size and effect size was approximately .38, well below the conventional .80 threshold, indicating that the study was underpowered to detect an effect of the observed magnitude.

Table 2

Prerequisite Tests and Inferential Hypothesis Testing Results

Variable / Test	Experimental (n = 29)	Control (n = 29)	Test Statistic	p-value	Decision
Posttest normality (Shapiro–Wilk)	W = 0.892	—	—	.008	Non-normal
Variance homogeneity (Levene)	—	—	F = 5.814	.019	Heteroscedastic
Posttest comparison (Mann–Whitney U)	Mean rank = 30.84	Mean rank = 28.16	U = 381.50; Z = −0.708	.479	Not significant
Effect size (Cohen's d)	—	—	d = 0.47 [−0.05, 0.99]	—	Small-to-moderate
Achieved power (post-hoc)	—	—	1 − β = 0.38	—	Underpowered

*Note. The 95% confidence interval for Cohen's d was computed using the non-central t distribution. Power was estimated with G*Power 3.1.9.7 for a two-tailed Mann–Whitney test with n = 29 per group, α = .05, and d = 0.47.*

The non-significant result must be interpreted in light of the descriptive pattern. The experimental group's higher N-Gain (0.6413 versus 0.4755) and higher posttest mean represent a descriptive advantage, but the inferential test cannot rule out chance variation at conventional thresholds. Because the confidence interval around d includes zero, the data do not establish a reliable achievement advantage for PBL. The correct conclusion is that the present study is inconclusive regarding between-group differences, not that PBL is ineffective.

Qualitative Observation Results

Structured observations across the three PBL sessions documented differences in verbal participation among students. During the first session, high-achieving students tended to lead rule identification, while some lower-achieving students copied answers without explaining their reasoning. In the second session, reduced verbal involvement was observed among some lower-achieving students during complex sentence parsing. By the third session, some high-achieving

students completed tasks independently, while other students participated less actively. These observations describe classroom interaction and reveal quantifiable trends across intervention sessions.

Table 3

Distribution of Verbal Participation Turns by Higher and Lower Achievement Groups Across Three PBL Sessions

Session	Topic	Total Turns (n)	High Achievers (%)	Low Achievers (%)	Dominance Index	Observed Behavior
1	Definiteness & gender	97	57.2%	10.3%	5.55	High achievers led rule identification; low achievers copied answers without verbalizing rationale.
2	Case endings (i'rāb)	88	64.8%	9.1%	7.12	Low achievers withdrew from discussion during complex sentence parsing.
3	Number consistency	83	68.4%	8.4%	8.14	Group polarization peaked; high achievers completed tasks independently, reducing interaction to passive consensus.

Note. Percentages are based on total coded verbal participation turns per session (Session 1: n = 97 turns; Session 2: n = 88 turns; Session 3: n = 83 turns). Higher (Top Quartile) and lower (Bottom Quartile) achievement groups were defined by quartile split on the AGAT pretest within the experimental class. Dominance index = high-achiever turns ÷ low-achiever turns.

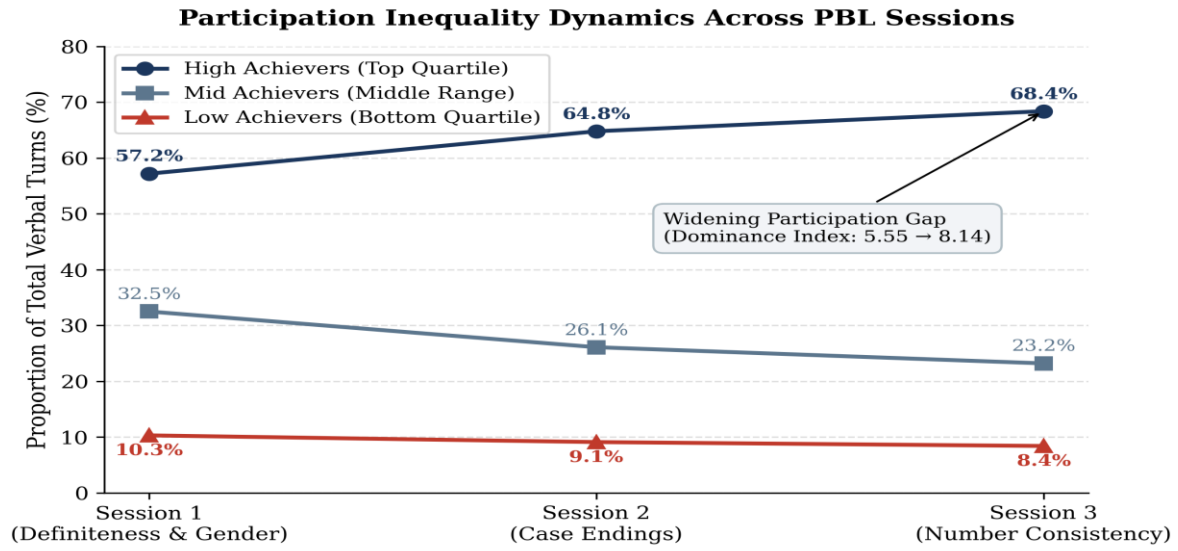


Figure 2

Line graph displaying proportion of total verbal turns (%) across three intervention sessions for high and low achievers

Note. Proportion of total verbal turns (%) plotted across three intervention sessions for high and low achievers in the experimental class.

Two features of these results merit emphasis before the Discussion. First, the participation data are descriptive and do not, by themselves, establish cognitive overload; they document a behavioral pattern of unequal engagement consistent with, but not proof of, differential cognitive burden. Second, we observed only the experimental class, so we have no parallel participation data for the lecture condition. Therefore, comparing participation dynamics across instructional methods remains a question for future research rather than a finding of this study.

Discussion

The Descriptive–Inferential Split and What It Does Not Mean

The study's central finding is the dissociation between descriptive improvement and inferential non-significance. The PBL class recorded a higher N-Gain (0.6413 versus 0.4755) and a higher posttest mean (79.28 versus 71.03), yet the Mann–Whitney U test yielded $p = .479$ with an effect size of $d = 0.47$ whose 95% confidence interval [−0.05, 0.99] crosses zero. The temptation to attribute this split solely to low statistical power should be resisted. Achieved power of .38 means the study was underpowered, but underpowered studies produce imprecise estimates, not automatically true effects hidden by sample size (Faul et al., 2007). The honest reading is that the data are compatible with a small negative effect and with a moderate positive effect, and the study cannot adjudicate between them. This statement reflects the limits of the evidence, not a claim that PBL works and the statistics failed to show it.

The data do support a narrower, more defensible conclusion. The experimental group's posttest standard deviation (20.26) was substantially larger than the control group's (13.92), a 46% difference in dispersion. Combined with the participation data reported in Table 3 and Figure 2, this dispersion suggests that the PBL condition produced divergent trajectories within the same class. Some students gained considerably, others gained little or regressed relative to peers. A treatment that raises the mean while widening the variance is not uniformly effective. Its benefits depend on something the design did not measure directly—most plausibly, prior knowledge and the ability to self-regulate inquiry without external structure.

Mechanism: Why Unguided Inquiry Meets Resistance in *Na'at–Man'ūt*

The *na'at–man'ūt* construction is not a rule to be recalled but a coordination task to be executed. Four dimensions—definiteness, case ending, gender, and number—must be satisfied simultaneously; violating any one invalidates the construction. In CLT terms, this is high element interactivity: students cannot process the elements serially because each element's correctness depends on the others (Sweller, 2020; Sweller et al., 2023). For a novice, coordinating four dimensions already approaches working-memory capacity before any problem-solving demand is added. PBL adds that demand by requiring students to discover the coordination pattern while simultaneously applying it to an ill-structured problem. The result is what the Procedural-Constructivist Conflict predicts: the inquiry process consumes the working-memory resources that schema construction requires.

This mechanism is consistent with the participation trajectory in Figure 2. If unguided inquiry's cognitive cost is borne unequally, students with stronger prior schemas should dominate the verbal space because they can afford the coordination cost and still have capacity left for articulation. The observed pattern—high achievers rising from 57.2% to 68.4% of turns while low achievers fell from 10.3% to 8.4%—fits this prediction. The widening dominance index across sessions (5.55 → 7.12 → 8.14) tracks the increasing element interactivity of the session topics, from definiteness and gender in Session 1 to number consistency in Session 3. This is a behavioral signature of differential load, not direct evidence of it. The study did not administer the cognitive load scale consistently, so overload remains a theoretical interpretation rather than a measured finding, and it should be read as such.

The mechanism also explains the control group's moderate N-Gain of 0.4755. Lecture-based instruction provides explicit rule explanations and worked examples, which, for novices, reduces extraneous load during early schema formation (Kirschner et al., 2006). The control group did not reach the experimental group's descriptive mean, but it also showed less dispersion. Explicit instruction appears to have produced a more uniform, if lower-ceiling, outcome. This reflects the trade-off the Procedural-Constructivist Conflict describes: guidance protects novices from overload but limits exploratory depth, while unguided inquiry offers depth but leaves novices behind.

Comparison with Prior Research

These findings align with recent experimental work on instructional sequencing. Zambrano et al. (2026) showed that for learners with low prior knowledge, explicit instruction before inquiry yields better outcomes than inquiry-first sequences, a pattern attributable to the expertise reversal effect. The present study extends that logic to Arabic grammar: the *na'at–man'ūt* domain is precisely the kind of high-interactivity content where inquiry-first sequencing should disadvantage novices most. This finding also aligns with Ngu et al. (2026), who showed that managing element interactivity through completion problems and worked examples works better for novices than unguided discovery. The control group's performance is consistent with that literature: explicit guidance did not produce greater gains, but it did produce more stable ones.

The participation findings resonate with research on group dynamics in active learning. Frambach et al. (2014) and Qutieshat et al. (2020) documented peer dominance and participation disparity in PBL groups across cultural and disciplinary contexts. The present data add a grammar-specific dimension to that literature: the disparity widened as the grammatical content became more interactive, suggesting that participation inequality in PBL is not merely a social phenomenon but is modulated by the cognitive demands of the task. This is a modest but genuine contribution, because it connects the participation literature to CLT in a domain—Arabic grammar—where that connection has not been empirically traced. The scaffolding literature offers a further point of contact. Hendrayana and Mutaqin (2025) found that scaffolding within PBL disproportionately benefited students with low and medium prior knowledge, precisely the students who withdrew in the present study. Their finding supports the inference that the

participation gap observed here is not an inevitable feature of PBL but a consequence of unscaffolded PBL. The present study did not test that inference, but it identifies the population—low-prior-knowledge novices in high-interactivity grammar—for whom scaffolding would be expected to matter most.

Theoretical and Practical Implications

This study's theoretical contribution is the empirical articulation of the Procedural-Constructivist Conflict. The conflict is not a new theory; it restates, in domain-specific terms, the long-standing tension between constructivist inquiry and cognitive load limits that Kirschner et al. (2006) analyzed. This study shows that the tension is not abstract but observable in a specific grammatical construction and classroom. The *na'at–man'ūt* system is a useful test case because its element interactivity is unambiguous and its failure modes are visible in student errors and in the withdrawal of low achievers from discussion. For practitioners in madrasah and *pesantren* contexts, the implication is cautionary rather than prohibitive. PBL should not be abandoned, but its unscaffolded application to high-interactivity grammar topics should be.

The participation data suggest that in a three-session PBL unit without structured support, the students who most need to practice the coordination task are the least likely to do so verbally. A teacher who observes high achievers dominating discussion should read that pattern as a signal that the task's cognitive demands are unevenly distributed, not as evidence that the low achievers are unmotivated. The practical response is to build in the supports that the four-phase model proposes—schema activation before inquiry, guided analysis before open problem-solving—rather than to remove the inquiry task altogether.

The Scaffolded-PBL Model as a Proposition, Not a Finding

The four-phase model presented in Table 4 is a post-hoc proposition derived from the observed participation patterns and the CLT literature. It was not tested in this study, and no claim of effectiveness is made for it. Its value is heuristic: it specifies a sequence—Schema Activation, Guided Analysis, Open Problem-Solving, Reflective Consolidation—that follows directly from the mechanism identified in the Discussion.

Schema activation addresses the prior-knowledge gap that left low achievers passive. Guided analysis reduces element interactivity by having students complete problems before full problem-solving is required. Open problem-solving preserves the constructivist goal once schema support is in place. Reflective consolidation targets the metacognitive dimension that unguided PBL tends to leave implicit. Each phase has a testable indicator: participation equity across achievement groups, error-type distribution in guided tasks, transfer performance on novel constructions, and self-explanation quality. Future research can evaluate these indicators; the present study cannot.

Table 4

Proposed Scaffolded-PBL Model: Phases, Procedures, and Cognitive Load Targets

Phase	Instructional Focus	Scaffolding Procedure	Cognitive Load Target	Testable Indicator
1. Schema Activation	Explicit rule orientation	Worked examples, visual alignment matrices, explicit rule modeling	Minimize extraneous load	Pretest-to-guided-task error reduction
2. Guided Analysis	Structured diagnosis	Completion problems, partially filled syntax trees, faded prompts	Support intrinsic load	Error-type distribution across achievement groups
3. Open Problem-Solving	Collaborative application	Authentic text error analysis, role-rotated peer diagnosis	Promote germane load	Participation equity index
4. Reflective Consolidation	Metacognitive synthesis	Self-explanation prompts, rule generalization, error reflection	Consolidate schemas	Self-explanation quality and transfer scores

Note. The model is a conceptual proposition derived from the present findings and the CLT literature. It was not empirically tested in this study. Indicators are proposed for future evaluation.

Limitations and Directions for Future Research

Four limitations bound the interpretation of this study. First, the sample of 58 students in two intact classes yields an achieved power of .38, which is insufficient for definitive hypothesis testing. The confidence interval around d is wide and includes zero; the study is inconclusive, and neither exonerates nor condemns PBL. Second, the absence of consistent cognitive load measurement means that overload is inferred from participation and dispersion rather than measured directly. Future studies should administer a validated scale, such as the Leppink et al. (2013) instrument or the MCLS-POL (Andersen & Makransky, 2021), at each session and pair it with a process measure, such as eye-tracking or think-aloud protocols, to triangulate subjective and objective load. Third, the three-session intervention is short; schema-building effects may require longer exposure, and participation may stabilize or reverse over a full unit. Fourth, the absence of a scaffolded-PBL condition means the model proposed in Table 4 remains untested. A four-arm design—lecture, unscaffolded PBL, scaffolded PBL, and scaffolded PBL with fading—would allow the Procedural-Constructivist Conflict to be tested rather than merely illustrated.

Future research should also measure participation equity directly. The dominance index used here is a descriptive statistic; a formal index with a known sampling distribution would allow researchers to test participation inequality as an outcome, not just report it as a pattern. Combining such an index with cognitive load measurement would permit a mediation analysis: does load mediate the relationship between instructional condition and participation equity?

That question naturally follows the present study, and it requires the larger sample and instrumented design this study could not provide.

CONCLUSION

This study compared Problem-Based Learning (PBL) with lecture-based instruction for teaching the *na'at–man'ūt* construction to eleventh-grade students in two intact classes at an Islamic senior high school in Indonesia. The PBL class recorded a higher descriptive N-Gain (0.6413 versus 0.4755) and a higher posttest mean, but the Mann–Whitney U test was not statistically significant (U = 381.50, p = .479), and the 95% confidence interval for the effect size (d = 0.47) spanned from −0.05 to 0.99. With achieved power of only .38, the study cannot adjudicate between a small negative effect and a moderate positive effect. The defensible conclusion, therefore, is that the present evidence is inconclusive regarding between-group achievement differences, not that PBL is ineffective. This distinction matters: an underpowered null result is an absence of evidence, and treating it as evidence of absence would misrepresent what the data can support.

The study's substantive contribution lies in the process data rather than the achievement comparison. Structured observations across three PBL sessions documented a widening participation gap: high-achieving students increased their share of verbal turns from 57.2% to 68.4%. In contrast, low-achieving students' share declined from 10.3% to 8.4%, and the dominance index rose from 5.55 to 8.14 as the element interactivity of the grammatical content increased. The experimental group's posttest standard deviation (20.26) was substantially larger than the control group's (13.92), indicating divergent trajectories within the PBL class. These patterns align with the Procedural-Constructivist Conflict—the proposition that unscaffolded inquiry imposes extraneous load on novices processing high-interactivity content, leaving those with weaker prior schemas unable to participate fully. However, they do not prove cognitive overload, because the cognitive load scale was not administered consistently and no direct load measurement is reported. The mechanism remains a theoretically grounded interpretation rather than a measured finding, and future work should treat it as a hypothesis to be tested.

The theoretical contribution is empirically specifying that conflict in a domain where element interactivity is unambiguous. The *na'at–man'ūt* system requires simultaneous coordination of definiteness, case, gender, and number, and the observed withdrawal of low achievers tracked the increasing coordination demands of successive session topics. This connects the participation literature on PBL group dynamics with cognitive load theory in Arabic grammar, a connection that had not been empirically traced. The four-phase Scaffolded-PBL model—Schema Activation, Guided Analysis, Open Problem-Solving, and Reflective Consolidation—is offered as a testable proposition derived from these findings, not as a validated intervention. Its phases specify indicators that a future study could evaluate: participation equity across achievement groups, error-type distribution in guided tasks, and transfer performance on novel constructions.

Three limitations bound these conclusions. The sample of 58 students in two intact classes limits statistical power and precludes causal inference, since class membership and treatment condition are confounded. The three-session intervention is short relative to the schema-building processes under study, and the participation trajectory may stabilize or reverse over a longer unit. Most importantly, the absence of a scaffolded-PBL condition means the proposed model remains untested; the study illustrates the conflict it names but does not evaluate the solution it proposes. Future research should adopt a four-arm design—lecture, unscaffolded PBL, scaffolded PBL, and scaffolded PBL with fading—with a sample size sufficient to detect the smallest meaningful effect, cognitive load measured in each session using a validated instrument, and a formal participation equity index amenable to mediation analysis.

ACKNOWLEDGMENT

All authors contributed substantially to this manuscript. Moh Sholeh Afyuddin contributed to conceptualization, methodology, field data collection, and drafting the initial manuscript. Muflihah Rahayu contributed to formal data analysis and critical revision. M. Syamsul Ma'arif contributed to literature review and data interpretation. Munjiyat contributed to theoretical framing and final proofreading. All authors have read and agreed to the published version of the manuscript.

DECLARATIONS

Declaration Category	Statement Details
Use of AI Statement	The author(s) declare(s) that no Artificial Intelligence (AI) tools were used in the writing or analytical process of this manuscript.
Authors' Contributions	All authors contributed substantially to this manuscript. Moh Sholeh Afyuddin contributed to conceptualization, methodology, field data collection, and drafting the initial manuscript. Muflihah Rahayu contributed to formal data analysis and critical revision. M. Syamsul Ma'arif contributed to literature review and data interpretation. Munjiyat contributed to theoretical framing and final proofreading. All authors have read and agreed to the published version of the manuscript.
Funding Statement	This research received no external funding.
Conflict of Interest	The author(s) declare(s) that there is no conflict of interest regarding the publication of this article.

REFERENCES

Abdurrohman, A., & Ulinnuhaa, M. (2024). Problematika pembelajaran *qawaid* bahasa Arab di madrasah aliyah: Analisis kesulitan dan solusi metodologis. *Jurnal Pendidikan Bahasa Arab*, 12(1), 45–58. https://doi.org/10.21043/jpba.v12i1.8892

Abu Yousef, S. S., & Abu Owda, M. F. (2026). The Effect of Interaction between Active Learning Strategies and Cognitive Style on Reducing EFL Learners' Reading Anxiety. *An-Najah University Journal for Research - B (Humanities), 40 (6), pp. 563 - 582.* https://doi.org/10.35552/0247.40.5.2566

Afiyah, N. (2025). Deklaratif ke prosedural: Otomatisasi aturan tata bahasa Arab melalui latihan terstruktur. *Lisanudrak*, 17(1), 112–128. https://doi.org/10.21111/lisan.v17i1.4421

Al-Busaidi, S., Yusuf, T., & Reinders, H. (2021). A model for implementing problem-based language learning: Exploring pedagogical affordances and teacher perceptions. *International Journal of Learning, Teaching and Educational Research*, 20(1), 1–22. https://doi.org/10.26803/ijlter.20.1.1

Allen, D. E., Donham, R. S., & Bernhardt, S. A. (2011). Problem-based learning. *New Directions for Teaching and Learning*, 2011(128), 21–29. https://doi.org/10.1002/tl.465

Andersen, M. S., & Makransky, G. (2021). The Validation and Further Development of the Multidimensional Cognitive Load Scale for Physical and Online Lectures (MCLS-POL). *Frontiers in Psychology, 12, art. no. 642084.* https://doi.org/10.3389/fpsyg.2021.642084

Belland, B. R., Walker, A. E., & Kim, N. J. (2017). A Bayesian Network Meta-Analysis to Synthesize the Influence of Contexts of Scaffolding Use on Cognitive Outcomes in STEM Education. *Review of Educational Research, 87 (6), pp. 1042 - 1081.* https://doi.org/10.3102/0034654317723009

Boulaares, D.M. (2026). From declarative knowledge to procedural fluency: a generative AI tutor for Arabic agreement rules. *Computer Assisted Language Learning.* https://doi.org/10.1080/09588221.2026.2685745

Bozorgian, H., Muhammadpour, M., & Alehchali, A. B. (2026). The effects of metacognitive intervention and metacognitive intervention through dialogic interaction on L2 learners' listening comprehension and cognitive load. *Language Awareness, 35 (1), pp. 271 - 297.* https://doi.org/10.1080/09658416.2025.2489937

Choo, S. S. Y. (2014). Scaffolding in problem-based learning. *One-Day, One-Problem: An Approach to Problem-based Learning, pp. 167 - 184.* https://doi.org/10.1007/978-981-4021-75-3_8

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

Creswell, J. W., & Plano Clark, V. L. (2018). *Designing and conducting mixed methods research* (3rd ed.). SAGE Publications.

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

Fauzi, M. (2021). Kecemasan dalam pembelajaran *qawaid* dan dampaknya terhadap kemahiran berbicara bahasa Arab. *Arabiyatuna: Jurnal Bahasa Arab*, 5(2), 211–228. https://doi.org/10.29240/jba.v5i2.2890

Frambach, J. M., Driessen, E. W., Beh, P., & van der Vleuten, C. P. M. (2014). Rethinking the globalization of problem-based learning: How culture challenges self-directed learning. *Medical Education*, 48(8), 738–750. https://doi.org/10.1111/medu.12484

Hendrayana, A., & Mutaqin, A. (2025). The effectiveness of problem-based learning through scaffolding on students' mathematical critical thinking and cognitive load. *Educational Process: International Journal*, 16(2), 145–162. https://doi.org/10.22521/edupij.2025.16.275

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

Hung, B. P., Vien, T., & Vu, N. N. (2018). Applying cognitive linguistics to teaching English prepositions: A quasi-experimental study. *International Journal of Instruction, 11 (3), pp. 327 - 346.* https://doi.org/10.12973/iji.2018.11323a

Kirschner, P. A., Sweller, J., & Clark, R. E. (2006). Why minimal guidance during instruction does not work: An analysis of constructivist, discovery, problem-based, experiential, and inquiry-based teaching. *Educational Psychologist*, 41(2), 75–86. https://doi.org/10.1207/s15326985ep4102_1

Krieglstein F., Beege M., Wesenberg L., Rey G.D., Schneider S. (2025). The Distorting Influence of Primacy Effects on Reporting Cognitive Load in Learning Materials of Varying Complexity. *Educational Psychology Review, 37 (1), art. no. 2.* https://doi.org/10.1007/s10648-024-09980-0

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

Lange, C., Costley, J. (2022). How sequencing content and fading instructional support reduce extraneous processing. *Innovations in Education and Teaching International, 59 (3), pp. 253 - 262.* https://doi.org/10.1080/14703297.2020.1861966

Leahy, W., & Sweller, J. (2019). Cognitive load theory, resource depletion, and the transient information effect. *Educational Psychology Review*, 31(1), 157–178. https://doi.org/10.1007/s10648-018-9461-8

Leppink, J., Paas, F., Van der Vleuten, C. P. M., Van Gog, T., & Van Merriënboer, J. J. G. (2013). Development of a cognitive load scale for problem solving and lecture-based tasks. *Contemporary Educational Psychology*, 38(1), 92–102. https://doi.org/10.1016/j.cedpsych.2012.12.001

Liu, L. (2026). Adaptive Learning Platforms in English Education: Enhancing Comprehensive Skills and Social Interaction in the Digital Classroom. *European Journal of Social Science Education and Research, 13 (1), pp. 203 - 225.* https://doi.org/10.26417/xscbj161

Makransky G., Terkildsen T.S., Mayer R.E. (2019). Role of subjective and objective measures of cognitive processing during learning in explaining the spatial contiguity effect. *Learning and Instruction, 61, pp. 23 - 34.* https://doi.org/10.1016/j.learninstruc.2018.12.001

Moeen A.A., Nejadansari D., Dabaghi A. (2019). The impact of implicit vs explicit grammar teaching through scaffolding on Iranian learners' speaking achievement: focusing on fluency, accuracy, and complexity. *Journal of Applied Research in Higher Education, 11 (4), pp. 800 - 813.* https://doi.org/10.1108/JARHE-01-2019-0021

Mohammed R., Zouhair G., & Noureddine B., Mohamed A.L.-H. (2025). Reading disabilities in Arabic-speaking learners: A comprehensive systematic review of screening tests and challenges ahead. *International Journal of Educational Research Open, 9, art. no. 100541.* https://doi.org/10.1016/j.ijedro.2025.100541

Moslemi Nezhad Arani S., Zarei A.A., Sarani A. (2023). Problem-based language learning: Why Aren't teachers using it?. *Social Sciences and Humanities Open, 8 (1), art. no. 100668.* https://doi.org/10.1016/j.ssaho.2023.100668

Ngu, B. H., Phan, H. P., & Usop, H. (2026). Optimizing instruction to reduce element interactivity in procedural learning tasks. *Journal of Educational Psychology*, 118(1), 88–104. https://doi.org/10.1037/edu0000891

Qutieshat, A. S., Abusamak, M. O., & Maragha, T. N. (2020). Peer dominance and participation inequality in group-based active learning. *Medical Teacher*, 42(6), 675–682. https://doi.org/10.1080/0142159X.2020.1732808

Simons, K. D., & Klein, J. D. (2007). The impact of scaffolding and student achievement levels in a problem-based learning environment. *Instructional Science, 35 (1), pp. 41 - 72.* https://doi.org/10.1002/acp.70039

Sweller, J. (2020). Cognitive load theory and educational technology. *Educational Technology Research and Development*, 68(1), 1–16. https://doi.org/10.1007/s11423-019-09701-3

Sweller, J., van Merriënboer, J. J. G., & Paas, F. (2023). A cognitive load theory view of active learning and direct instruction. *Educational Psychology Review*, 35(1), Article 8. https://doi.org/10.1007/s10648-023-09722-2

Zambrano R., J., Narváez-Rivera, C., Sayago-Heredia, G., & Jácome-León, S. (2026). The expertise reversal effect in inquiry-based language learning: An experimental study. *Learning and Instruction*, 98, Article 101912. https://doi.org/10.1016/j.learninstruc.2025.101912

Zhang L., Saisai Z. (2026). Intelligent personal assistants as 'scaffolding on demand': Mechanisms of impact on L2 listening performance and anxiety among EFL learners. *Acta Psychologica, 263, art. no. 106246.* https://doi.org/10.1016/j.actpsy.2026.106246

de Jong, T. (2010). Cognitive load theory, educational research, and instructional design: Some food for thought. *Instructional Science*, 38(2), 105–134. https://doi.org/10.1007/s11251-009-9110-0