Volume 11 Number 2, August 2026, pp. 209-221
P-ISSN:2477-7935
E-ISSN: 2548-6225
DOI: 10.59052/edufisika.v11i2.56090

A 3x3 LATIN SQUARE COUNTERBALANCED DESIGN FOR COMPARING INSTRUCTIONAL TREATMENTS IN PHYSICS EDUCATION

Febriana Eka Damayanti[1], Iwan Permana Suwarna[2]
Physics Education, Faculty of Tarbiyah and Teacher Training, UIN Syarif Hidayatullah Jakarta
Corresponding author email: iwan.permana.suwarna@uinjkt.ac.id

Article Info
Received: 25 Mei 2026
Accepted: 18 Juni 2026
Publication: 1 August 2026

Abstract :
Conventional quasi-experimental designs in physics education are prone to bias due to differences in characteristics between classes, order-of-instruction effects, and the influence of prior treatments. This study introduces and demonstrates a balanced 3×3 Latin Square design as an alternative methodological framework to enhance internal validity in physics education experiments based on in Indonesia. Using a quasi-experimental rotation at SMAN 4 South Tangerang (N = 102; three 11th-grade classes), each class received three instructional treatments in rotation: (A) Problem-Based Learning (PBL) aided by Zep Quiz, (B) Zep Quiz without PBL, and (C) conventional lectures, on three subtopics of Dynamics of Motion. Students' critical thinking skills were assessed using an instrument based on Facione's six indicators. Data were analyzed using prerequisite tests (Shapiro-Wilk, Levene, Mauchly), Repeated-Measures ANOVA, Latin Square ANOVA, and a carryover effect test. Three main findings emerged: (1) a significant period effect ($p = 0.028$, $\eta^2 = 0.035$) indicated a gradual increase in scores over time; (2) the class factor accounted for the majority of the variance ($F = 153.051$, $p < 0.001$, $\eta^2 = 0.736$), far exceeding the effect of the learning method ($F = 0.678$, $p = 0.058$, $\eta^2 = 0.005$); and (3) the carryover effect was not significant ($F = 0.000$, $p = 1.000$), confirming the effectiveness of rotation in minimizing cross-treatment contamination. These findings establish the 3×3 Latin square design as a practical and bias-controlled framework for research on physics education in whole-class settings.

Keywords: Critical Thinking; Counterbalanced Design; Latin Square; Physics Learning

This is open access article under the CC BY-SA licence

INTRODUCTION

The learning of physics necessitates that students acquire the ability of critical thinking as an essential competence to reason and solve problems in the current learning setting. From the international study done using PISA 2022, the average science results among Indonesian students are below the average of the OECD (383 vs. 485) (OECD, 2023). The same trend has been shown through studies

done in South Tangerang where it is observed that the physics higher-order thinking skill among the students is low, whereby the average scores obtained are much lower than the midpoint of the assessment scale, hence showing that the situation is not just in the national assessment but also in the classroom level (Zulfiani et al., 2020).

Critical thinking in this study is understood as a collection of mindful, self-regulatory judgement skills, including interpretation, analysis, evaluation, inference, and explanation, which was operationalized based on Facione's Delphi consensus report criteria (1990) that has become widely used for formal educational assessment structures. Critical thinking, on the other hand, focuses on scientific reasoning and complex problem-solving rather than simply mastering concepts (Fadly et al., 2025). Research on critical thinking skills in Indonesian physics classrooms continues to grow, with studies consistently documenting low baseline performance and supporting the effectiveness of active-learning approaches such as Problem-Based Learning (Arni, 2025; Sari et al., 2025). More broadly, systematic reviews of critical thinking instruction confirm that structured pedagogical interventions yield moderate positive effects across educational levels (Tiruneh et al., 2014; Abrami et al., 2015).

Claims made regarding the effectiveness of instructional treatment methods require just as much emphasis on good research design as on the method of instruction itself. The existing research on methodological issues has stressed the importance of the fact that both measured and unmeasured confounders can affect associations, and the need for using a design that includes proper controls is crucial for minimizing bias and establishing causation (D'Onofrio et al., 2020). Before determining which instructional treatment method works better than another, there must be the ability to isolate the treatment method's effect from alternative explanations.

The traditional between-group quasi-experiments, especially pretest-posttest nonequivalent control group design, can suffer from problems of internal validity since intact classes may differ systematically with respect to prior achievement, classroom atmosphere and teaching styles. In sequential or counterbalanced experiments, there can also be some carryover effect because of the exposure to a particular treatment in the previous period(s), which biases the estimation of the treatment effect if not modeled properly (Ho & Min, 2025). These problems become even more serious in the case of packaged treatments such as "PBL assisted by a digital medium," as the observed changes cannot be isolated to the learning approach, the medium or the interaction between them without a proper design.

Latin square design can offer a practical solution for the above problems because each treatment will appear only once in each row and in each column of the design matrix; therefore, all groups will experience all the treatments at the same time while controlling for two sources of extraneous variability (class and period) (Sule et al., 2025).

Prior physics education research in Indonesia employing quasi-experimental designs has typically assigned fixed treatment-control class pairings, with each class receiving only one instructional approach (Anantasia & Rindrayani, 2025). This structural choice introduces at least three compounded methodological vulnerabilities. First, pre-existing class differences in prior achievement, motivation, and classroom climate are treated as random error rather than as estimable variance components, so that post-treatment differences between classes cannot be cleanly attributed to the instructional treatment (D'Onofrio et al., 2020). Second, when learning models are bundled with digital media platforms, between-group designs cannot disentangle the contribution of the pedagogical model from that of the medium (Rusmana et al., 2025). Third, order effects and carryover contamination are rarely modeled or tested in Indonesian physics education research, despite representing plausible alternative explanations for observed learning gains in any multi-period study (Ho & Min, 2025; Kulnik et al., 2025). These compounded weaknesses motivate the adoption of a design that simultaneously addresses all three threats.

This study thus seeks to achieve the following three aims: (1) to show if it is feasible to employ 3×3 Latin Square counterbalancing design under real-life conditions of Indonesian high schools; (2) to decompose outcome variability in terms of method, period, and class factors employing Latin Square ANOVA; and (3) to perform an analysis of carryover contamination to confirm that the designed rotation and interperiod washout time were sufficient to reduce the impact of previous treatments. This study will rotate (A) PBL with Zep Quiz, (B) Zep Quiz without PBL, and (C) traditional lecturing with respect to three subtopics of Motion Dynamics in order to evaluate critical thinking skills of students with the

help of the Facione scale. To the best of our knowledge, this is the first study in Indonesian classroom physics education to employ all three techniques in one realization.

RESEARCH METHOD

This section describes the methodological implementation of a 3×3 Latin Square counterbalanced design, organized into four subsections: research design, participants and setting, research procedure, and data analysis strategy. The research utilized a quasi-experimental design, which is one of the most frequently used approaches in educational research, particularly when random assignment cannot be applied (Creswell & Creswell, 2018). The research took place in the high school SMAN 4 Tangerang Selatan for the even semester of 2025/2026 school year, and its subject matter was Motion Dynamics, which included three sub-topics.

Research Design

This study employed a quasi-experimental counterbalanced approach structured as a 3×3 Latin Square to compare three instructional treatments while simultaneously controlling period and carryover effects. The three treatments were: (A) Problem-Based Learning (PBL) assisted by Zep Quiz, (B) Zep Quiz without PBL, and (C) conventional lecture. Each intact class received all three treatments exactly once, but in systematically varied orders across three learning periods, each corresponding to one Motion Dynamics subtopic (Table 1).

Three primary reasons motivated the use of 3×3 Latin Square Design as opposed to others. As opposed to the pre-test post-test nonequivalent control group design, the Latin Square avoids any possibility of between-groups confound because all groups are presented with all treatments so that no pre-existing group characteristics can bias any of the conditions. As opposed to a repeated measures design that does not utilize counterbalancing, the Latin Square provides a way of balancing treatment order over classes and periods so that any potential effects of order are evenly distributed. As opposed to complex factorial designs or randomized block designs that require individual participant level randomization, which is structurally impossible in Indonesian high schools' intact classes, the Latin Square achieves similar partitioning of variances through intact classes. Modern methodological literature still supports the use of Latin Square designs as valid approaches for controlling multiple sources of nuisance factors.

Table 1. Treatment Rotation Pattern Based on Latin Square 3×3

Class/Period	Period 1 (Newton's Laws)	Period 2 (Momentum & Impulse)	Period 3 (Rotational Motion)
XI.1	A: PBL + Zep Quiz	B: Zep Quiz	C: Lecture
XI.2	B: Zep Quiz	C: Lecture	A: PBL + Zep Quiz
XI.3	C: Lecture	A: PBL + Zep Quiz	B: Zep Quiz

The teaching method is the key independent variable that is going to be tested for its effect; it forms the matrix cells. Class and learning period are nuisance variables since they have an impact on critical thinking performance but are not of major research interest. Including them as matrix rows and columns allows for estimating their effects and separating them from the teaching method effect. However, if the teaching method becomes a nuisance variable, the effect of interest is going to become incomprehensible (Kirk, 2013; Montgomery, 2017; Richardson, 2018).

Latin square design is preferred among other designs because of the following three strengths. Relative to a quasi-experiment which assigns one treatment to each group, the Latin square design avoids any bias that may exist between the classes since each class will receive all treatments. Relative to simple repeated measures designs, which lack counterbalancing, the Latin square design ensures that period effects are evenly distributed among all groups at all periods. Relative to factorial designs and completely randomised block designs, which entail individual-level random assignment and are difficult to carry out in intact classes in schools, the Latin square design can distinguish multiple variances (Sule

et al., 2025; Lim et al., 2021; Richardson, 2018; Kirk, 2013). The Latin Square methodology was proven to be effective again in recent literature on methodology in ensuring the reliability of results obtained after controlling several nuisance variables at the same time without losing the interpretation capability of the statistical analysis (Livingston & Rayner, 2024). No other methodology can accomplish this disentanglement together with carryover effect testing in the same analysis.

Selection of the three treatment methods was guided by three related aspects. First, PBL was chosen because it has been consistently demonstrated to improve critical thinking through authentic, student-centred problem-solving (Arni, 2025; Tiruneh et al., 2017). Its integration with Zep Quiz aimed to leverage instant technology-mediated feedback to support the metacognitive cycle in PBL, reflecting on understanding and adjusting thinking strategies independently (Bellhäuser et al., 2023). Second, Zep Quiz without PBL was included as a standalone condition to isolate the contribution of the digital medium from that of the pedagogical approach. Without this condition, it would be impossible to determine whether any improvement in critical thinking originated from the PBL structure, the instant feedback of Zep Quiz, or their synergy, following the component-control principle in simplified factorial designs (Montgomery, 2017). Third, conventional lecture was included as a baseline comparison condition reflecting the most common instructional practice in Indonesian secondary schools, making the findings directly relevant to practitioners considering the adoption of active learning methods (Anantasia & Rindrayani, 2025).

Concerning the five-day washout period between interventions, this decision was made on the basis of two factors. First, methodologically speaking, a certain washout period should be provided for the effect of the previous intervention to fade out before applying the next one, which is recommended in crossover experiments for behavioral and educational research. (Lim & In, 2021; Kulnik et al., 2025). While cognitive carryover can hardly be prevented because of the accumulative character of learning, an appropriate interval at least allows reducing the novelty effect and cross-method contamination. Second, in terms of practical issues, the academic schedule of the school does not allow for longer washout periods without affecting the continuity of the course. A five-day washout period (one week of the school year) was chosen as a compromise between splitting subtopics (Newtons Laws, Momentum and Impulse, and Rotational Motion), according to the syllabus, and giving enough time to reduce cross-contamination. This fact is proven by the insignificant result of the carryover test ($F = 0.000$; $p = 1.000$).

Research Participants and Setting

The subjects used in this study were three complete groups of Grade XI MIPA students of SMAN 4 Tangerang Selatan ($N = 102$; three groups with 34 students each) in even semester of the year 2025/2026. The three groups (XI.1, XI.2, and XI.3) served as the accessible subjects for Latin Square design under identical curricula and similar scheduling conditions. The teaching was done by one physics teacher only in order to reduce teacher variable influence.

Equivalence at the baseline among classes was checked by means of the Kruskal-Wallis test before applying the treatment; the results can be seen in the Results section. The primary measure of effect is the level of critical thinking skill among the students, measured after each phase of the experiment. It should be mentioned that while equivalence at the baseline proves the lack of any differences in the initial level of the students' skills before the treatment, it cannot prove that the process of studying in the classrooms was identical in all other respects throughout the experiment. Therefore, the class is included into the Latin Square design (Kirk, 2013; Montgomery, 2017).

Instrument Development and Validation

The entire research methodology process is illustrated in Figure 1 below.

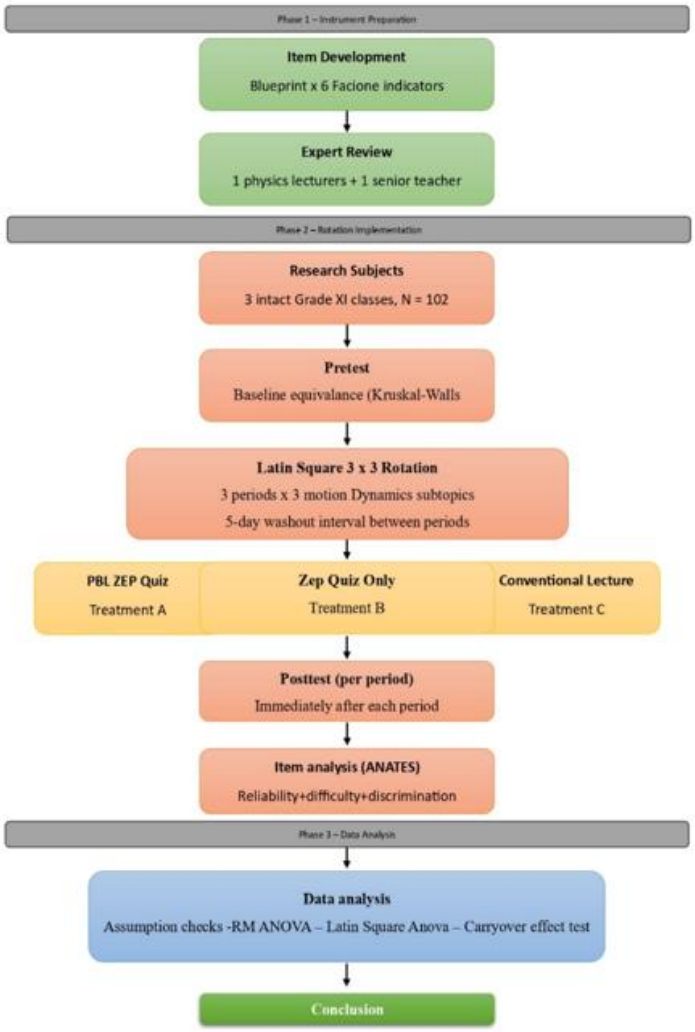

Figure 1. Research Framework for the 3×3 Latin Square Counterbalanced Design

The test for critical thinking skills involved three groups of multiplechoice questions on different topics, all based on the six criteria set out by Facione, namely: interpretation, analysis, evaluation, inference, explanation, and self-regulation. The instrument development process consisted of two phases.

Stage 1 (development of the blueprint and expert evaluation) involved the development of the items based on a blueprint that mapped each question according to a critical thinking skill and its associated learning objective. The content validity of the test was evaluated by two experts: a physics education lecturer from UIN Syarif Hidayatullah Jakarta and a senior physics teacher. Any item considered not appropriate and unclear was modified and/or eliminated. In Stage 2 (test items quality analysis), the characteristics of the instrument were analyzed using Classical Test Theory employing ANATES Version 4.0 according to Field (2018). These analyses included: (a) validity through item-total score correlation (items without significant correlations were excluded); (b) indices of item difficulty; (c) discrimination indices; and (d) reliability coefficients.

Item difficulty index distribution is shown in Table 2 below. Most items under Newton's Laws and Momentum & Impulse fall within the easy to moderate category, suggesting that they are appropriate for the average ability of Grade XI students. In case of Rotational Motion, there was more variation in difficulty levels, with six items rated as difficult to very difficult.

Table 2. Distribution of Item Difficulty Levels by Topic

Difficulty Category	Newton's Law	Momentum&Impuls	Rotational Motion
Very Easy	12	16	8
Easy	4	1	2
Moderate	2	1	2
Difficult	0	0	2
Very Difficult	0	0	4
Total items	18	18	18

The reliability was computed using the Spearman-Brown formula, with results as follows: Reliability = 0.79 for Newton's Laws; Reliability = 0.78 for Momentum and Impulse; Reliability = 0.45 for Rotational Motion. The reliability figures for the first two tests were found to be satisfactory. The lower reliability score for Rotational Motion test can be explained by reduced instructional time (one class against two for others) and inclusion of higher level questions leading to reduced reliability figure.

Research Procedure

This research process comprised four steps. (1) Preparation: Materials for learning, lesson plans, and Zep Quiz settings were prepared for each treatment and subject. (2) Pre-test: The critical thinking ability of students was tested before starting the treatment process to ensure that baseline equivalence is achieved. (3) Rotation: Three treatments were conducted in three periods as stated in Table 1, where posttest was conducted at the end of each period with a five-day washout period between consecutive periods. (4) Data analysis: Pre-test and post-test scores were recorded, period gain scores were calculated, and data analysis was ready for inferential statistics.

Data Analysis Technique

Analysis of data involved (i) assumption verification, (ii) evaluation of changes that took place over time periods, (iii) breakdown of variance into factors of method, period, and class in the context of the Latin Square design, and (iv) carryover effect test. Assumptions verification included testing for normality through Shapiro-Wilk, homogeneity of variances through Levene's test, and sphericity through Mauchly's test; each test was assessed based on established criteria for parametric tests (Blanca et al., 2023). Changes over time periods were examined using repeated measures ANOVA. Then, Latin Square ANOVA was employed to assess and breakdown variance into method, period, and class factors according to the methodology suggested by Montgomery (2017); the magnitude of effect is measured as partial eta squared (η^2). The presence of carryover effect was tested by considering the previous period treatment as a factor; if the effect is non-significant ($p > 0.05$), then there is no contamination of the following periods by previous treatments. Analysis of statistical data was conducted with the help of IBM SPSS Statistics Version 31.

Table 3. Stages of Data Analysis Techniques

Analysis Stage	Purpose	Statistical Test / Model	Decision Criteria	Result in This Study
Assumption checks	Verify analysis prerequisites	Shapiro-Wilk; Levene; Mauchly	Obtain information on data distribution and variance homogeneity	Shapiro-Wilk: $p = 0.072$; 0.125; 0.058 (mild non-normality, not extreme); Levene: all $p > 0.05$, variances homogeneous; Mauchly: $W = 0.982$, $p > 0.05$ (sphericity met, RM ANOVA robust under these conditions per Blanca et al., 2023)

Baseline equivalence	Check pre-rotation comparability	Kruskal–Wallis	$p > 0.05$ indicates comparable baseline	Newton's Laws: H=5.511, p=0.064; Momentum & Impulse: H=0.024, p=0.988; Rotational Motion: H=2.690, p=0.261. All equivalent.
Period effect	Examine change across periods	Repeated-measures ANOVA	$p < 0.05$ indicates period-related change	$F = 3.633$; $p = 0.028$; $\eta^2 = 0.035$ (significant, small effect)
Latin Square decomposition	Separate method, period, and class effects	Latin Square ANOVA	F, p for Method/Period/Class	Method: F=0.678, p=0.058, η^2=0.005; Period: F=3.128, p=0.045, η^2=0.020; Class: F=153.051, p<0.001, η^2=0.736
Effect magnitude	Quantify practical influence	Partial eta squared (η^2)	Report η^2 for key effects	Class dominates (73.6%); method trivial (0.5%); period small (2.0%)
Carryover test	Detect prior-treatment contamination	Carryover effect	$p > 0.05$ indicates no carryover	$F = 0.000$; $p = 1.000$; $\eta^2 = 0.000$. No carryover detected.

Checks for assumption violations were done in sequence before conducting inferential tests. The first assumption test involved the use of Shapiro-Wilk normality tests to get score data from each treatment cell. Results of these tests revealed slight violations of normality in some cells but not severely. As suggested by Blanca et al. (2023), repeated-measures ANOVA provides robust and reliable output, especially if there is a slight violation of the normality assumption as long as the experiment is balanced and the groups are equal in size, with epsilon values closer to 1.0. The second assumption test was conducted to confirm that there was homogeneity of within-group variances ($p > 0.05$) in treatment conditions. This satisfied another necessary condition for the use of parametric test in the analysis. Thirdly, Mauchly's sphericity test, which is the unique prerequisite of Repeated-Measures ANOVA, was done to establish the violation of the sphericity assumption, which was found to be violated ($W = 0.982$; $p > 0.05$).

The choice of the Kruskal-Wallis test for verifying baseline equivalence was justified based on the outcome of assumption checking since there were slight violations of the assumptions of normality for some pretest data, thus providing a more reliable alternative for comparison of three independent samples before treatment than the F-test (Field, 2018). The lack of significant between-class difference in critical thinking ability was established for all topics, which provided a proper basis for further comparisons.

Following the checking of baseline equivalence and parametric assumptions for post-treatment data, the analysis was performed. Repeated-Measures ANOVA revealed changes in critical thinking through three learning periods. Then, Latin Square ANOVA identified factors of influence of teaching method, learning period, and between-class characteristics on the variance in scores. The impact of each of these factors was measured using partial eta squared. Finally, carryover effect test ensured that washout periods were long enough to get rid of the possible carryover effect of previous treatment.

RESULTS AND DISCUSSION

This section will state the results produced by the 3×3 Latin Square counterbalanced design and interpret them from a methodology point of view. In the Results section, there are tests that check baseline equivalency, assumptions and the Latin Square variance analysis. In the Discussion section, an

explanation is provided of what these results say about the advantages and drawbacks of using a counterbalanced approach for obtaining internal validity in physics education studies.

Result

The critical thinking abilities of the three intact classes were statistically equal before the rotation began. Tests on each subtopic revealed that there were no significant differences between classes (Newton's Laws: H = 5.511, p = 0.064; Momentum and Impulse: H = 0.024, p = 0.988; Rotational Motion: H = 2.690, p = 0.261), thus ensuring internal validity of further comparisons (Table 4).

Table 4. Results of Initial Equivalence Test (Kruskal-Wallis)

Topic	H	df	p	Description
Newton's Laws	5.511	2	0.064	Equivalent (p > 0.05)
Momentum & Impulse	0.024	2	0.988	Equivalent (p > 0.05)
Rotational Motion	2.690	2	0.261	Equivalent (p > 0.05)

Tests using Shapiro-Wilk showed that all gain scores were not normally distributed in all the conditions. Levene test results showed that the variances were relatively homogenous in all conditions except one. Mauchly's test results were used to show that the sphericity assumption holds (W = .982, p > 0.05). This allowed for ANOVA with repeated measures and Latin Square designs to be used without correction.

Descriptive statistics showed that all three instructional treatments were associated with improved post-instruction performance (Table 4). Zep Quiz yielded the highest mean posttest score (M = 68.2, SD = 16.8), followed by PBL + Zep Quiz (M = 65.8, SD = 14.5) and Lecture (M = 62.2, SD = 17.7). These values are interpreted as performance tendencies only; the inferential analyses presented below indicate that the method factor did not reach statistical significance, and therefore these descriptive differences should not be interpreted as evidence of differential effectiveness.

Table 5. Descriptive Statistics of Posttest Scores by Treatment

Statistic	ZepQuiz	PBL+Zep	Lecture
Mean	68.2	65.8	62.2
SD	16.8	14.5	17.7
Min	28	33	16
Max	94	94	94

Students' critical-thinking performance differed significantly across learning periods (F (2, 202) = 3.633, p = 0.028, η^2 = 0.035), indicating a period-related change regardless of treatment. This small-to-moderate effect implies that scores improved over the sequence of learning periods, but that the period contribution to overall variance was limited in magnitude.

Latin Square ANOVA decomposed outcome variance into three orthogonal components. The method factor did not reach statistical significance (F (2, 299) = 0.678, p = 0.058, η^2 = 0.005), indicating that between-treatment differences were not sufficiently large relative to within-design variability to be distinguished under this sample size. The period factor was significant (F (2, 299) = 3.128, p = 0.045, η^2 = 0.020). In regard to the class factor, there was a highly significant impact (F (2, 299) = 153.051, p < 0.001, η^2 = 0.736), indicating that variation between classes was the prevailing factor that contributed to variability in this data set, contributing around 73.6% of the variance, which is much greater than that of instructional technique (Table 6).

Table 6. Results of Repeated Measures ANOVA, Latin Square ANOVA, and Carryover Test

Analysis	Source of Variance	SS	df	MS	F	p	η^2	Interpretation

RM ANOVA	Time (Between Periods)	2541.595	2	1270.797	3.633	0.028	0.035	Significant, small effect
	Error	70654.405	202	349.774	-	-	-	
Latin Square ANOVA	Method	183.948	2	91.974	0.678	0.058	0.005	Non-significant, negligible effect
	Period	848.359	2	424.180	3.128	0.045	0.020	Significant, small effect
	Class	41508.908	2	20754.454	153.051	<0.001	0.736	Significant, very large effect
	Error	40545.840	299	135.605	-	-	-	
Carryover Effect		0.000	2	0.000	0.000	1.000	0.000	Non-significant, no carryover

Source: IBM SPSS Statistics Version 31 output, 2026

Results from the carryover test indicate that exposure to the treatment in previous periods does not affect the outcomes in the following period ($F = 0.000$, $p = 1.000$, $\eta^2 = 0.000$). The finding is a proof of the validity of the rotation design and the five-day washout period.

Discussion

The first significant methodology used in this study is the application of the 3x3 Latin square design in which there is a balanced exposure to three instruction treatments across both classes and periods at the same time. The Latin square designs have been recommended as means to control extraneous factors of variability in education research (Sule et al., 2025). This study provides an illustration of how this recommendation can be implemented in the classroom setting in Indonesia.

The class effect ($F = 153.051$, $\eta^2 = 0.736$) has proven to be the most important result of this research. It should be noted that an η^2 value of 0.736 means that inter-class difference accounted for about 73.6% of total variation in Latin Square design, being approximately 147 times greater than the variance due to instructional approach ($\eta^2 = 0.005$). This discrepancy has significant methodological implications, since the use of a simple between groups quasi-experimental design for comparison of any two of these classes under different conditions would attribute most of the class differences to the effect of treatment, thus giving an inflated estimate of the instructional superiority.

There are several ways through which this large variance might have arisen. First, classroom learning environments and norms in regards to peer interactions differ between intact classes and cannot be captured through simple equivalence testing on critical thinking scores. Classes with a collaborative learning environment will benefit much more from PBL and interactive media, while those with a passive learning approach will gain equally from direct instruction. Second, differences in previous exposure to technology among students could have made the students differentially ready for Zep Quiz, and the pre-test, which was only measuring critical thinking, could not capture that factor. Third, even with one teacher in charge, slight differences in the rate and student responsiveness among the three classes may have increased the effect of the ability difference. Previous research that has been conducted using a multilevel model shows that teacher/class effects can be as great as school effects in secondary education, which proves that class effects are legitimate and substantial sources of variance (Stallasch et al., 2021). Most importantly, the Latin square design is able to measure this effect instead of including it in error.

Since the result is consistent across the two different statistical analyses – namely, the Latin square and the ANCOVA analyses carried out on the identical instructional dataset, this provides further assurance that variance due to classes was a real context effect and not an artefact of a specific statistical method of analysis.

Non-significant method factor ($p = 0.058$, $\eta^2 = 0.005$) implies that differences between treatments could not be detected statistically under the current design and sample size. This outcome is confirmed by the related study conducted on the same dataset through ANCOVA and pretest scores as the covariate that revealed the non-significant effect of the model ($F (2, 202) = 2.229$, $p = 0.109$, $\eta^2 =$

0.015) meaning that the variations of instructional model explained very little variance of outcomes when controlled for the initial ability level (Luthfiyani & Suwarna, 2026). The concurrence of the findings obtained via Latin Square ANOVA and ANCOVA leads to the conclusion about the comparative effectiveness of all three treatments. In this case, the descriptive advantage of Zep Quiz (M = 68.2 vs. 65.8 and 62.2) cannot be considered as an indication of superiority since it means that it represented the inconsistent performance pattern that could not be detected statistically. Theoretically, Zep Quiz's minor descriptive advantage can be viewed as a result of the proved positive effects of immediate feedback rich in information for self-regulated learning (Bellhäuser et al., 2023) and good interactive representations according to multimedia learning theory (Mayer, 2024). This is also in line with findings that analytical reasoning in physics involves interaction with multiple representations (Docktor & Mestre, 2014). But adding PBL syntax with interactive media can lead to high extraneous cognitive load unless scaffolding is appropriately managed, possibly explaining why the addition of PBL was not more beneficial than Zep Quiz alone (Chen et al., 2023).

The failure to find a significant effect of the method is not an indication that the methods were equivalent in their educational value. In fact, this is an indication that whatever the effects were due to the treatments were rather minor in comparison to the variance attributed to class and period variations.

This is because of the existence of significant period effect ($p = 0.028$, $\eta^2 = 0.035$), which implies that there was some change in student performance over different learning periods, due to learning, practice, or the level of difficulty of the subtopic. Period effects are common in repeated measures situations and have been found to act as threats which may be misinterpreted as treatment effects if there is no control on order. Counterbalancing is the recommended solution to this problem (Newsom, n.d.). In the Latin square design, treatments are balanced over different periods, but the remaining period effect indicates that there is a difference in the level of difficulty of the three subtopics or increased familiarity with the Facione-based test.

The nonsignificance of the carryover effect ($F = 0.000$, $p = 1.000$) is methodologically justified by the fact that previous treatment did not have a systematic contaminating influence on subsequent results in this implementation. In the field of education, the complete avoidance of carryover is impossible due to its cumulative nature; the nonsignificance of the carryover effect must thus be understood as a demonstration of the fact that previous instructional format had no additional influence apart from the learning conditions of the present period (Kulnik et al., 2025; Lim et al.,2021).

Limitations

There are three limitations of this study. First, the sample is limited to only three classes in one school and thus does not have the capability of being generalized to other Indonesian educational contexts. The small sample size of 102 participants limited statistical power in detecting a small difference in methods because the between-class variance was extremely large compared to the other methods. Future research is encouraged to do an a priori power analysis and use a more sizable sample. Second, the Rotational Motion instrument had a low reliability ($r = 0.45$) due to the less instruction time given (only one meeting versus two for other topics) and hard items that compressed the distribution, resulting in a low reliability coefficient that decreased measurement precision of critical thinking for that topic. Third, treatment fidelity was not properly checked since only one instructor conducted all the lessons; differences between the classes in implementation and classroom management can potentially become factors affecting the small difference in the methods.

CONCLUSION

This This study proves that 3×3 Latin Square design with counterbalancing is indeed feasible and meaningful methodologically in terms of physics education research on intact classes in Indonesia. With its capability to distinguish between the influences of teaching method, learning period, and differences between classes, and to detect any carryover effects from previous treatments, this design offers a better base for causal inference than traditional treatment comparison designs.

Three important findings can be made based on the analysis conducted. Firstly, the critical thinking ability of students improved systematically throughout learning periods in all cases regardless of method used, albeit with a small effect size ($\eta^2 = 0.035$), which means that the accumulation of learning processes and different levels of topic complexity somewhat influenced test score increases.

Secondly, there was no difference in effectiveness among the three teaching methods ($\eta^2 = 0.005$), while differences between classes played the dominant role in students' critical thinking achievements ($\eta^2 = 0.736$), which underscores the problem of misinterpretation in traditional designs in which between-class variance is inevitably intertwined with treatment effects. Finally, there was no evidence of a carryover effect ($p = 1.000$).

The major contribution of this research is based on the empirical side of the experiment; the Latin Square design can be applied in realistic conditions of Indonesian schools, and the hidden structure of critical thinking variation can be revealed due to the dominant role of between-class variance. The above results imply that both the choice of research design and an instructional approach are equally significant when trying to prove something in the field of physics education. Future researches should use data about classes and students' characteristics, employ multilevel models in order to identify the sources of between-class variance, perform a priori power analysis, and check for rotation effects.

ACKNOWLEDGMENTS

We hereby express our deep appreciation to the principal, teachers, and students of SMAN 4 South Tangerang for the cooperation shown in participating in this research. We would also like to thank the Physics Education Department of UIN Syarif Hidayatullah Jakarta for all the academic assistance and guidance received in this research. The authors further thank Fikriyah Luthpiyani for her assistance, support, and collaboration throughout the research process and in completing this article.

REFERENCES

Abrami, P. C., Bernard, R. M., Borokhovski, E., Waddington, D. I., Wade, C. A., & Persson, T. (2015). Strategies for teaching students to think critically: A meta-analysis. Review of Educational Research, 85(2), 275–314. https://doi.org/10.3102/0034654314551063

Afriani, L., Mutmainnah, & Sunarni. (2025). Understanding the design of research and development methods in the field of education. Jurnal Pendidikan dan Pengembangan, 6(1), 1–5.

Anantasia, G., & Rindrayani, S. R. (2025). Metodologi penelitian quasi eksperimen. ADIBA: Journal of Education, 5(2), 183–192.

Arni, K. J. (2025). Systematic review of research trends on critical thinking skills in physics learning. International Journal of Science Education and Science, 2(1), 43–52. https://doi.org/10.56566/ijses.v2i1.262

Bellhäuser, H., Dignath, C., & Theobald, M. (2023). Daily automated feedback enhances self-regulated learning: A longitudinal randomized field experiment. Frontiers in Psychology, 14, Article 1125873. https://doi.org/10.3389/fpsyg.2023.1125873

Blanca, M. J., Alarcón, R., Arnau, J., Bono, R., & Bendayan, R. (2023). Repeated measures ANOVA and adjusted F-tests when sphericity is violated: Which procedure is best? Frontiers in Psychology, 14, 1192453. https://doi.org/10.3389/fpsyg.2023.1192453

Chen, O., Paas, F., & Sweller, J. (2023). A cognitive load theory approach to defining and measuring task complexity through element interactivity. Educational Psychology Review, 35, Article 63. https://doi.org/10.1007/s10648-023-09782-w

Creswell, J. W., & Creswell, J. D. (2018). Research design: Qualitative, quantitative, and mixed methods approaches (5th ed.). SAGE Publications.

D'Onofrio, B. M., Sjölander, A., Lahey, B. B., Lichtenstein, P., & Öberg, A. S. (2020). Accounting for confounding in observational studies. Annual Review of Clinical Psychology, 16, 25–48. https://doi.org/10.1146/annurev-clinpsy-032816-045030

Docktor, J. L., & Mestre, J. P. (2014). Synthesis of discipline-based education research in physics. Physical Review Special Topics-Physics Education Research, 10(2), 020119. https://doi.org/10.1103/PhysRevSTPER.10.020119

Facione, P. A. (1990). Critical thinking: A statement of expert consensus for purposes of educational assessment and instruction (The Delphi Report) (ERIC Document No. ED315423). American Philosophical Association.

Fadly, M., Simatupang, H. N., Simanjorang, A. S., Purba, J. M., Mihardi, S., & Bunawan, W. (2025). Analisis kemampuan berpikir kritis siswa pada pelajaran fisika tingkat menengah atas. SAKOLA - Journal of Sains Cooperative Learning and Law, 2(1), 434-436.

Field, A. (2018). Discovering statistics using IBM SPSS statistics (5th ed.). SAGE Publications.

Ho, J., & Min, J. (2025). Causal inference in counterbalanced within-subjects designs. arXiv. (https://arxiv.org/abs/2505.03937)

Kirk, R. E. (2013). Experimental design: Procedures for the behavioral sciences (4th ed.). SAGE Publications.

Kulnik, S. T., Carrozzo, A. E., & Crutzen, R. (2025). Persistent carry-over in a two-period randomised crossover design for behavioural interventions without the expectation of return to baseline after intervention cessation. Trials, 26, 561. https://doi.org/10.1186/s13063-025-09358-9

Lim, C.-Y., & In, J. (2021). Considerations for crossover design in clinical study. Korean Journal of Anesthesiology, 74(4), 293–299. https://doi.org/10.4097/kja.21165

Livingston Jr, G. C., & Rayner, J. C. W. (2024). Rank tests for the latin square design. *Communications in Statistics - Theory and Methods*, 1–14. https://doi.org/10.1080/03610926.2024.2387249

Luthfiyani, F., & Suwarna, I. P. (2026). A counterbalanced quasi-experiment comparing Problem-Based Learning with Zep Quiz, Zep Quiz alone, and traditional lecture on critical thinking in high school mechanics. Jurnal Penelitian Ilmu Pendidikan, 19(1), 23-36. (https://doi.org/10.21831/jpip.v19i1.97460)

Mayer, R. E. (2024). The past, present, and future of the cognitive theory of multimedia learning. Educational Psychology Review, 36, Article 8. https://doi.org/10.1007/s10648-023-09842-1

Montgomery, D. C. (2017). Design and analysis of experiments (9th ed.). John Wiley & Sons.

Newsom, J. T. (n.d.). Precautions with within-subjects experimental designs [PDF handout]. Portland State University. https://web.pdx.edu/~newsomj/uvclass/ho_counterbalancing.pdf

OECD. (2023). PISA 2022 results (Volume I): The state of learning and equity in education. OECD Publishing. https://doi.org/10.1787/53f23881-en

Parker, K., Nunns, M., Xiao, Z., Ford, T., Stallard, P., Kuyken, W., Axford, N., & Ukoumunne, O. C. (2025). Patterns of intra-cluster correlation coefficients in school-based cluster randomised controlled trials of interventions for improving social-emotional functioning outcomes in pupils: A secondary data analysis of five UK-based studies. BMC Medical Research Methodology, 25, Article 120. (https://doi.org/10.1186/s12874-025-02574-6)

Richardson, J. T. E. (2018). The use of Latin-square designs in educational and psychological research. Educational Research Review, 24, 84–97. https://doi.org/10.1016/j.edurev.2018.03.002

Rusmana, A., Maulidya, R. A., Awalludin, M. R., & Mulyanti, D. (2025). Pengaruh penggunaan model pembelajaran PjBL dengan bantuan media pembelajaran Zep Quis terhadap kemampuan berpikir kritis dan kreatif siswa sekolah dasar kelas tinggi. Journal of Elementary Education, 8(4), 774–780.

Santi, L. M. (2020). Implementasi pembelajaran matematika dengan model problem-based learning di sekolah menengah pertama (kajian meta-analisis) (Skripsi Sarjana, Universitas Pendidikan Indonesia).

Sari, S., Zuhdi, M., & Taufik, M. (2025) Pengaruh model problem-based learning berbantuan media digital terhadap kemampuan berpikir kritis dan hasil belajar fisika. Jurnal Ilmiah Profesi Pendidikan, 10(3), 2234–2238. https://doi.org/10.29303/jipp.v10i3.3798.

Stallasch, S. E., Lüdtke, O., Artelt, C., & Brunner, M. (2021). Multilevel design parameters to plan cluster-randomized intervention studies on student achievement in elementary and secondary school. Journal of Research on Educational Effectiveness, 14(1), 172–206. https://doi.org/10.1080/19345747.2020.1823539

Sule, A. S., Lantana, B., Nuhu, U. M., Nasir, U., & Amina, A. N. (2025). Latin square design: A methodological innovation in educational research for improved teaching and learning. FUDMA Journal of Sciences, 9(5), 328–334. https://doi.org/10.33003/fjs-2025-0905-3686

Tiruneh, D. T., De Cock, M., Weldeslassie, A. G., Elen, J., & Janssen, R. (2017). Measuring critical thinking in physics: Development and validation of a critical thinking test in electricity and magnetism. International Journal of Science and Mathematics Education, 15(4), 663–682. https://doi.org/10.1007/s10763-016-9723-0

Tiruneh, D. T., Verburgh, A., & Elen, J. (2014). Effectiveness of critical thinking instruction in higher education: A systematic review of intervention studies. Higher Education Studies, 4(1), 1–17. https://doi.org/10.5539/hes.v4n1p1

Yehezkiel, G., & Sapta, A. (2025). Efektivitas media pembelajaran digital terhadap minat dan motivasi belajar siswa di SMP (Systematic Literature Review). Jurnal Teknologi Pendidikan dan Pembelajaran (JTPP), 3(2), 386–395. https://jurnal.kopusindo.com/index.php/jtpp/index.

Zhang, L. (2025). Estimating school and teacher effects on students' academic performance using multilevel models with three or more levels: A literature review. Discover Education, 4, Article 4. https://doi.org/10.1007/s44217-024-00392-4

Zulfiani, Z., Suwarna, I. P., & Sumantri, M. F. (2020). Science adaptive assessment tool: Kolb's learning style profile and student's higher order thinking skill level. *Jurnal Pendidikan IPA Indonesia, 9*(2), 194–207. https://doi.org/10.15294/jpii.v9i2.23840