



The International Journal of Mathematics and Sciences Education

Publication details, including instructions for authors and subscription information:

<https://nakiscience.com/index.php/IJMSEd>



The International Journal of
Mathematics and Sciences Education



Designing a microsoft teams-based blended learning tool through design-based research to foster senior high school students' mathematical creative thinking in linear programming

**Erza Candra¹, Yumiati Yumiati²,
Endang Wahyuningrum³**

¹Department of Mathematics
Education, Universitas Terbuka,
Jakarta, Indonesia,
erzacandra7@gmail.com

²Department of Mathematics
Education, Universitas Terbuka,
Jakarta, Indonesia,
yumi@ecampus.ut.ac.id

³Department of Mathematics Education,
Universitas Terbuka, Jakarta, Indonesia,
endagw@ecampus.ut.ac.id

To cite this article: Candra, E., Yumiati, Y., & Wahyuningrum, E. (2026). Designing a microsoft teams-based blended learning tool through design-based research to foster senior high school students' mathematical creative thinking in linear programming. *International Journal of Mathematics and Sciences Education*, 4(2), 84-109.

To link to this article:

<https://nakiscience.com/index.php/IJMSEd>

Published by:

Nasir Al-Kutub Indonesia

Residential Street Kila Rengganis, Block I, Number 11, Labuapi, Indonesia, 83361

Designing a microsoft teams-based blended learning tool through design-based research to foster senior high school students' mathematical creative thinking in linear programming

Erza Candra^{1*}, Yumiati Yumiati², Endang Wahyuningrum³

¹Department of Mathematics Education, Universitas Terbuka, Jakarta, Indonesia, erzacandra7@gmail.com

²Department of Mathematics Education, Universitas Terbuka, Jakarta, Indonesia, yumi@ecampus.ut.ac.id

³Department of Mathematics Education, Universitas Terbuka, Jakarta, Indonesia, endagw@ecampus.ut.ac.id

*Corresponding Author: erzacandra7@gmail.com

Abstract

Mathematical creative thinking remains underdeveloped in Indonesian secondary classrooms: only 31% of Indonesian students reached baseline proficiency in PISA 2022, compared with 78% across OECD countries (OECD, 2024). In Linear Programming, instruction often emphasises procedures over flexible, original problem solving. This Design-Based Research (DBR) study developed and evaluated a Microsoft Teams-based blended learning toolkit to strengthen Grade XI science students' mathematical creative thinking. One DBR macro-cycle comprised four phases, Analysis and Design, Development, Implementation, and Evaluation, with two embedded design-test-revise micro-cycles. Needs analysis, expert review, and classroom observation informed development, while a quasi-experimental pretest-posttest evaluation involved 60 students in experimental and control groups at a private senior high school in Batam, Indonesia. Content and media validity ratings were 82% and 90%, respectively. However, the creative-thinking component scored 56%, prompting targeted redesign. Teacher and student responses indicated high practicality (83.33%). ANCOVA showed a higher adjusted posttest mean for the experimental group (75.52 versus 66.34; $F(1, 57) = 55.13$, $p < .001$, partial $\eta^2 = .49$). Normalised gain was moderate in the experimental group (0.33) and low in the control group (0.08). The findings suggest that iterative, technology-mediated design was associated with improved mathematical creative thinking and revealed design tensions requiring revision. Replication across schools is needed before generalising the resulting design principles.

Article History

Received:

14 August 2026

Revised:

28 September 2026

Accepted:

06 October 2026

Published Online:

07 October 2026

Keywords:

Blended learning;
Design-based research;
Linear programming
Microsoft Teams;
Mathematical creative thinking

1. Introduction

Creative thinking is now widely treated as a core 21st-century competency that formal assessment systems explicitly try to capture, and mathematics classrooms are expected to contribute directly to its development (OECD, [2023](#)). Within mathematics education specifically, creativity has long been theorised as the capacity to generate fluent, flexible, and original solution paths rather than to reproduce a single memorised procedure (Leikin & Pitta-Pantazi, [2013](#)). Recent empirical work continues to confirm that this capacity does not emerge automatically from conventional instruction: it must be deliberately designed for, whether through problem-posing tasks in algebra (Wilkie, [2024](#)), multiple-solution tasks whose creative demand can be traced through students' own solution processes (Schindler & Lilienthal, [2020](#)), or structured creative problem-solving interventions (Ersoy & Özcan Kara, [2026](#)).

At the national level, the PISA 2022 creative-thinking assessment provides direct evidence of a learner-performance gap: Indonesian 15-year-olds obtained a mean score of 19 out of 60, against an OECD average of 33, and only 31% reached the baseline proficiency level (Level 3), compared with 78% across OECD countries (OECD, [2024](#)). At the classroom level, this gap is visible in how students engage with contextual, open-ended mathematics problems. Studies of Linear Programming and related linear-inequality topics repeatedly report that students default to a single memorised graphical or substitution procedure and struggle to produce alternative models, flexible strategies, or genuinely novel interpretations of a constrained-optimisation scenario (Sari & Afriansyah, [2022](#); Wijaya et al., [2022](#)). This is a missed opportunity, because Linear Programming is structurally well suited to creative-thinking instruction: a single contextual optimisation problem can be modelled with different variable choices, solved through different constraint arrangements, and interpreted through different real-world framings (Sugiarto et al., [2021](#); Rahman et al., [2021](#)). Yet without a deliberately designed learning environment, this structural richness is rarely exploited in day-to-day teaching. This learner-performance gap is the first of three gaps that frame the present study.

Blended learning is frequently proposed as a pedagogical response to this kind of instructional gap, because it deliberately combines the structure and social presence of face-to-face teaching with the flexibility, self-pacing, and resource richness of an online environment (Garrison & Kanuka, [2004](#)). A growing international evidence base supports its potential: systematic mapping of the blended/hybrid learning literature documents its rapid scholarly expansion (Bozkurt, [2022](#)), while meta-analyses report generally positive, though heterogeneous, effects on performance, attitude, and engagement across countries (Cao, [2023](#)) and, more specifically, on mathematics achievement (Putri et al., [2025](#)). Emerging evidence even suggests that artificial-intelligence-enhanced blended designs can amplify these gains further (Wu et al., [2025](#)). At the same time, reviewers caution that the online component of blended learning is not automatically beneficial: its effectiveness depends heavily on how deliberately the online space is designed, scaffolded, and integrated with face-to-face activity (Rasheed et al., [2020](#)). Mathematics-specific studies echo this conditionality, showing that blended designs can raise both

achievement and behavioural engagement with mathematics, but only when the technology component is purposefully aligned with the pedagogical goal rather than treated as a mere content-delivery channel (Alamri, 2021; Attard & Holmes, 2020). What remains less clear, and constitutes a second, blended-learning gap, is how the online component should be designed when the target outcome is mathematical creative thinking rather than general achievement.

Microsoft Teams is one of several platforms that schools and universities have used for operationalising this kind of blended design in secondary and tertiary classrooms, particularly since institutions needed a single environment that could support synchronous video instruction, asynchronous assignment workflows, and structured file-based resource delivery (Situmorang, 2020). Local Indonesian studies report favourable student interest and learning outcomes when Microsoft Teams is used in mathematics instruction (Irmayanti et al., 2023), and larger-scale international studies confirm that, alongside tools such as Moodle and Zoom, it can meaningfully shape student engagement and achievement when institutions implement it strategically (Khrisat & Fakhouri, 2024). However, most of this Microsoft-Teams-specific mathematics literature evaluates a largely fixed platform configuration rather than treating the instructional design itself as an object of iterative, evidence-driven refinement.

This last point exposes the third, and most specific, gap addressed here: a design-methodological one. Much instructional-design research and development in Indonesian mathematics education is still reported as a largely linear, single-pass build-and-validate pipeline: a product is designed, checked once against expert validity and user practicality criteria, tested once for effectiveness, and reported, with evaluation functioning mainly as a final checkpoint rather than a genuine trigger for redesign. Design-Based Research (DBR) offers a theoretically more demanding alternative, in which each cycle's evaluation is treated as data that should, in principle, change the next cycle's design, and in which the goal is not only a validated artefact but also transferable design principles about how, when, and why the intervention works (The Design-Based Research Collective, 2003; Barab & Squire, 2004; Wang & Hannafin, 2005). DBR has an established international track record in K-12 and teacher-education research (Anderson & Shattuck, 2012), yet systematic reviews of the DBR literature show that mathematics education remains one of its comparatively under-represented application domains relative to science and general ICT-integration studies (Tinoca et al., 2022; Zheng, 2015), and comparable iterative, evidence-driven logic has proven productive when applied to instructional-design models more broadly (Spatioti et al., 2022).

Taken together, these three strands define the specific knowledge gap addressed by this study: little is known about how a Microsoft-Teams-based blended learning environment for Linear Programming can be iteratively designed, through evidence-driven redesign, so that its tasks and online components explicitly target the four indicators of mathematical creative thinking. Accordingly, the objective of this study was to design, develop, implement, and evaluate such a toolkit (a teaching module, student worksheets, instructional materials, and a mathematical creative-thinking test) for Grade

XI science students within a DBR framework. The intended contribution, which is distinct from the empirical findings reported in Section 3, is twofold: a documented case of how expert and user evaluation data were used as explicit redesign triggers in an Indonesian school setting, and a set of provisional, context-bound design principles (Section 3.2) that other teachers and designers can test in their own settings. Whether these principles transfer beyond the present school is a question for future iterations rather than a claim of this study.

A terminological clarification is needed before the research questions are stated. In much of the DBR literature, a cycle comprises design, enactment, analysis, and redesign (The Design-Based Research Collective, 2003; Wang & Hannafin, 2005). The four stages used in this study (Analysis and Design, Development, Implementation, and Evaluation) are therefore treated as phases of a single DBR macro-cycle rather than as four separate cycles. Within this macro-cycle, two complete micro-cycles of design, testing, analysis, and redesign were conducted: (a) expert validation followed by targeted revision of the materials, and (b) a limited try-out with 15 students followed by further revision (see Section 2.1). The Implementation and Evaluation phases constitute the enactment and retrospective analysis of the macro-cycle; a second macro-cycle, in which the revised design would be re-enacted, lies outside the scope of this article. Four research questions, one per phase, guided the study:

RQ1 (Analysis and Design phase). What learning needs, curricular demands, and platform constraints should inform the design principles of a Microsoft-Teams-based blended learning toolkit intended to foster mathematical creative thinking in Linear Programming?

RQ2 (Development phase). To what extent is the resulting toolkit judged valid by content, media, and language experts, and what design tensions surface during iterative expert review and revision?

RQ3 (Implementation phase). How practical is the toolkit, as perceived by teachers and students, when enacted in an authentic blended classroom setting?

RQ4 (Evaluation phase). (a) Is the toolkit associated with greater improvement in students' mathematical creative thinking than conventional instruction, and (b) what does this evidence suggest about the underlying design principles?

RQ4(a) is answered with quantitative evidence: pretest and posttest scores on the creative-thinking test, analysed with ANCOVA (posttest adjusted for pretest), normalised gain, and standardised effect sizes with confidence intervals.

Two bodies of substantive theory, and one body of methodological theory, anchor the design reported in this article. The first concerns what blended, technology-mediated instruction is and how it should be structured. Blended learning is defined here, following its most widely cited theoretical formulation, as the thoughtful integration of face-to-face and online learning experiences, in which the online component is not an add-on but a genuine redesign of how, when, and where learning activity happens (Garrison & Kanuka, 2004). This definitional emphasis on thoughtful integration rather than mere technology addition is consistent with more recent systematic mapping of the field, which shows the

blended-learning literature increasingly focused on design conditions rather than on technology adoption as such (Bozkurt, [2022](#)). In this study, Microsoft Teams was treated as the specific technological substrate through which that integration was operationalised: its assignment workflow, file repository, and meeting functionality were mapped onto specific pedagogical functions (structured practice, resource delivery, and synchronous discussion, respectively) rather than used as an undifferentiated communication channel.

The second body of theory concerns what mathematical creative thinking is and how tasks can be engineered to elicit it. Mathematical creative thinking is defined here as the capacity to generate fluent (multiple), flexible (strategically varied), original (unconventional but valid), and elaborated (well-justified) responses to a mathematical task, a four-indicator framework with a long history in mathematics-education research (Leikin & Pitta-Pantazi, [2013](#)) and continued empirical support in recent studies of algebraic problem-posing (Wilkie, [2024](#)), multiple-solution tasks (Schindler & Lilienthal, [2020](#)), and structured creative-problem-solving instruction (Ersoy & Özcan Kara, [2026](#)). A recurring theme across this literature is that these four indicators are not incidental by-products of good teaching in general; they must be specifically engineered into task design, because a task that has only one correct solution path cannot, by construction, elicit fluency or flexibility regardless of how well it is otherwise taught. This theoretical point directly motivated the Development-phase finding reported in Section 3.3, in which the toolkit's overall content validity (82%) coexisted with a much lower validity score (56%) on the sub-aspect specifically assessing alignment with these four indicators — a finding that, in hindsight, is exactly what this theoretical framework would predict for materials that had not yet been explicitly audited against it.

The third, methodological body of theory — Design-Based Research — governs how this study treats the relationship between the first two. DBR is defined as an interventionist methodology combining theory-driven design of a learning environment with systematic, cyclical empirical evaluation, whose purpose is not only a validated product but also transferable design knowledge about how and why the intervention works (The Design-Based Research Collective, [2003](#)). Unlike a conventional single-pass development pipeline, in which evaluation functions mainly as a final quality check, DBR treats each cycle's evaluation as data that should, in principle, change the next cycle's design (Barab & Squire, [2004](#); Wang & Hannafin, [2005](#)). This distinction is what justifies the Results structure adopted in this article (Section 3): rather than reporting Analysis and Design, Development, Implementation, and Evaluation findings as four independent, sequentially ordered outcomes, each subsection explicitly traces how evidence from one phase changed a decision in the next, following reporting conventions increasingly recommended for DBR studies in K-12 and mathematics-specific contexts (Anderson & Shattuck, [2012](#); Tinoca et al., [2022](#); Zheng, [2015](#)).

2. Method

This study adopted a Design-Based Research approach from the outset, understood as an interventionist methodology that blends theory-driven design of a learning environment with systematic empirical evaluation, organised into iterative cycles of design, enactment, analysis, and redesign (The Design-Based Research Collective, 2003). Consistent with common reporting practice in DBR studies situated in K-12 mathematics settings (Tinoca et al., 2022), the study was organised into one DBR macro-cycle comprising four phases — Analysis and Design, Development, Implementation, and Evaluation — in which each phase's evaluation evidence was treated as an explicit input into the redesign of the next, rather than as a terminal report (Barab & Squire, 2004; Wang & Hannafin, 2005). The Evaluation phase additionally incorporated a quasi-experimental, pretest-posttest, nonequivalent control-group design to generate comparative evidence of the toolkit's effect on students' mathematical creative thinking.

2.1 Research Design

Design and setting. The study was conducted at a private senior high school (SMA Kartini) in Batam, Riau Islands, Indonesia, during the 2022/2023 academic year. Two intact Grade XI science classes were assigned, respectively, to an experimental condition that received the Microsoft-Teams-based blended learning toolkit and a control condition that received conventional face-to-face instruction on the same Linear Programming content and the same allocated time. Because the classes were pre-existing administrative groups, random assignment of individual students was not feasible; this constraint, common to school-based DBR and R&D studies, is addressed directly in the Limitations subsection of the Discussion.

Allocation of classes. Two intact Grade XI science classes were selected: XI MIPA 2 served as the experimental class and XI MIPA 3 as the control class. Selection and allocation were conducted at the class level rather than by randomly assigning individual students. Accordingly, comparability cannot be inferred from class labels or a nonsignificant pretest difference alone. Pretest scores provide the baseline for describing initial differences and, where the assumptions are met, for adjusting the posttest comparison through ANCOVA. The non-random allocation remains a limitation because adjustment can account for observed baseline scores but cannot remove unmeasured differences between the classes.

Control condition. The control class received conventional face-to-face instruction on the same Linear Programming content and with the same allocated instructional time as the experimental class. The comparator should be described in terms of the teaching activities and learning materials actually used, including teacher explanation, worked examples, practice tasks, feedback, and homework where these are documented. Any access to Microsoft Teams, digital materials, or open-ended tasks in the control class must also be recorded. The comparison is interpreted as an evaluation of the blended learning toolkit as a package, since the platform, task design, and organisation of learning may jointly contribute to the observed outcomes.

Figure 1

Four phases of the DBR macro-cycle and feedback for redesign

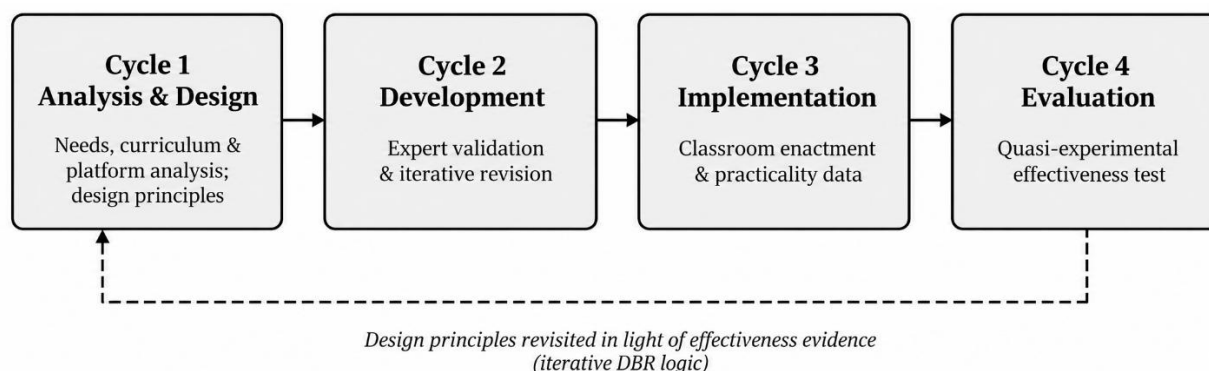


Figure 1 summarises the four phases of the DBR macro-cycle. During Development, the content-expert rating of 56% for alignment with creative-thinking indicators served as a trigger for revising the first prototype. The limited try-out with 15 students provided a subsequent opportunity to examine usability and identify wording problems before classroom implementation. At Evaluation, evidence from classroom use and the comparative assessment was considered in relation to the initial design principles. This feedback supports further redesign; it does not, by itself, demonstrate that a second macro-cycle was implemented. Figure 2 presents the sequence of the five blended sessions and the integration of face-to-face and Microsoft Teams components.

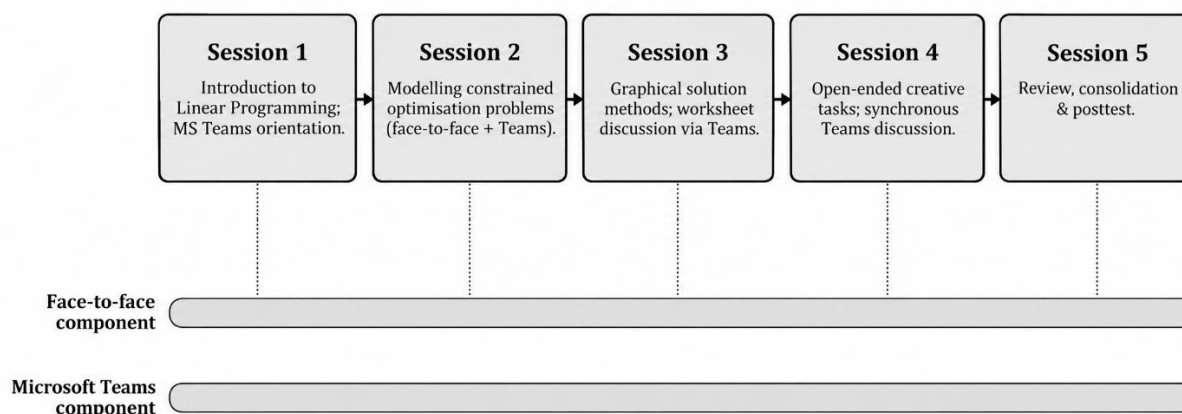
Sequence of iteration. The study took place during the 2022/2023 academic year. Analysis and Design preceded the development of the first prototype. Expert validation then identified insufficient alignment with creative-thinking indicators, leading to revision of the materials as documented in Table 3. The revised toolkit was subsequently examined in a limited try-out with 15 students, and the wording of two items was refined before implementation. The five-session classroom implementation was followed by evaluation of practicality and learning outcomes. Exact calendar dates and any formal expert re-rating of the revised prototype should be reported only when supported by the development records. The available account describes one macro-cycle, with formative testing and revision during development; no second classroom enactment of the final redesign is established.

Session structure. The toolkit was implemented across five blended sessions. The pretest preceded instruction in both classes. Session 1 introduced Linear Programming and oriented the experimental class to Microsoft Teams. Session 2 focused on modelling constrained optimisation problems; Session 3 addressed graphical solution methods and worksheet discussion through Teams; Session 4 used open-ended creative tasks and synchronous Teams discussion; and Session 5 provided review, consolidation, and the posttest. The sequence follows Figure 2. The face-to-face and Teams components were organised as parts of the same learning sequence. Their duration, frequency, and distribution between classroom and independent work require the timetable and

implementation records; no numerical estimate of platform use is inferred from the number of sessions alone.

Figure 2

Flow of the five blended implementation sessions



2.2 Participants

Participants. The Evaluation phase involved 60 Grade XI science students, with 30 students in the experimental class and 30 in the control class. Participants were recruited as intact classes. The reported class sizes do not establish the number of students with complete paired pretest and posttest records; the analysed sample should therefore be identified from the attendance and score records. The Analysis phase involved the mathematics teacher through interviews and observation, and the Development phase involved four expert validators: two content or materials experts, one media expert, and one language expert. The first author designed the toolkit as the teacher-researcher. Teaching responsibilities, test administration, test scoring, and questionnaire participation should be distinguished so that the reader can assess possible implementer and assessor bias. Age and gender statistics are omitted because no verified demographic data are supplied.

Prior to the main classroom trial, a limited try-out involved 15 students of varying mathematical ability (high, moderate, and low, as judged by the teacher) drawn from a class not involved in the main Evaluation phase. This limited try-out was deliberately kept separate from the 60 students in the Evaluation phase so that early exposure to the toolkit's items would not bias the subsequent pretest-posttest comparison, a precaution consistent with standard practice in instrument-development research more broadly.

2.3 Data Collection

Data collection. Consistent with DBR's characteristic reliance on mixed, multi-instrument data collection (Tinoca et al., 2022), four techniques were used across phases. (1) Semi-structured interviews and classroom observation informed the Analysis phase. (2) Expert validation questionnaires, covering content feasibility, presentation, alignment with creative-thinking indicators, instructional design, visual and technical media quality, and language readability, generated quantitative validity scores during the Development

phase. (3) Teacher and student response questionnaires, administered after the limited try-out and again after the main implementation, generated practicality data. (4) A mathematical creative-thinking test, consisting of open-ended items addressing the four canonical creative-thinking indicators (fluency, flexibility, originality, and elaboration), was administered as a pretest before instruction and a posttest after instruction to both the experimental and control classes.

Creative-thinking test. The assessment used open-ended Linear Programming tasks to examine fluency, flexibility, originality, and elaboration. Fluency concerns the production of multiple relevant ideas or answers; flexibility concerns the use of different valid approaches; originality concerns a mathematically valid response that shows an uncommon approach; and elaboration concerns the detail and justification of the solution. The number of items, the mapping of items to indicators, and the scoring scale must be taken from the administered instrument. Scores can be converted to a 0–100 scale using $S = (\text{obtained score} / \text{maximum possible score}) \times 100$, provided that the same scoring basis is applied to both classes and both assessments. Appendix A offers a proposed analytic rubric for checking the coherence of these indicators; it is not presented as the rubric used in the study. Rescaling the same model is not automatically a distinct idea, and the corner-point method and an iso-profit-line approach should count as different strategies only when the response demonstrates both correctly.

Instrument validity and reliability. Expert validation provided evidence concerning the suitability of the toolkit and its instructional materials. The reported percentage ratings do not, on their own, establish the content validity or reliability of the creative-thinking test and practicality questionnaires. The present documentation does not provide item-level Aiken's V coefficients, Cronbach's alpha estimates, or inter-rater agreement statistics; consequently, no numerical values are reported for these properties. The limited try-out with 15 students can inform clarity and usability, but cannot be treated as evidence of satisfactory reliability without the relevant response data and calculations. The number and identity of scorers, the rubric used, and the treatment of ambiguous responses should be documented. If scoring was conducted by a single assessor, the absence of an independent agreement estimate should be acknowledged as a limitation.

2.4 Data Analysis

Data analysis. Qualitative evidence from teacher interviews, classroom observation, and expert comments informed the identification of design needs and revision priorities. A thematic organisation of these materials, informed by Braun and Clarke (2006), should link each interpretation to its source and to the resulting design decision. The form of interview records, the coding responsibilities, and any review by another analyst require confirmation from the research records; audio recording, verbatim transcription, and consensus coding are not assumed. Validation and practicality scores were expressed as percentages using $P = (\text{sum of obtained scores} / \text{sum of maximum possible scores}) \times 100$. The denominator must reflect the actual instrument, response scale, and number of raters or respondents. An overall pooled percentage is appropriate only where the raw totals

are available; an unweighted average of aspect percentages should not be described as a pooled total. Numerical percentages should be retained, while category labels must follow one verified classification scheme consistently across Tables 2 and 4. For the quantitative evaluation, class-specific pretest and posttest means and standard deviations describe baseline performance and outcomes. An ANCOVA of posttest scores, with class as the factor and pretest as the covariate, is an appropriate proposed baseline-adjusted comparison for this nonequivalent-group design (Dimitrov & Rumrill, 2003), provided that the available data support the model. Before interpretation, examine the pretest–posttest relationship, overlap in baseline scores, residual distributions and variance, influential observations, and the class-by-pretest interaction to assess homogeneity of regression slopes. Shapiro–Wilk and Levene tests may supplement diagnostic plots but do not by themselves establish all ANCOVA assumptions. If ANCOVA has not been conducted, it must remain a proposed analysis rather than a reported procedure. Independent-samples comparisons of posttest and gain scores and within-class paired comparisons can provide supplementary evidence when their respective assumptions are met. Because the intervention was allocated to only one class per condition, the treatment is confounded with class, and student-level analyses cannot separate a toolkit effect from a class effect; causal conclusions must therefore remain cautious.

For an unadjusted comparison between the two classes, Cohen's d can be calculated as the difference between class means divided by the pooled standard deviation, $SD_{pooled} = \sqrt{\{(n_1 - 1)SD_1^2 + (n_2 - 1)SD_2^2\} / (n_1 + n_2 - 2)}$. The outcome used for d must be stated explicitly, and this estimate should not be described as an ANCOVA-adjusted effect. The conventional benchmarks of approximately 0.2, 0.5, and 0.8 provide descriptive guides to small, medium, and large differences (Cohen, 1988). Partial η^2 may be reported when ANCOVA is actually conducted; confidence intervals require the corresponding estimates and calculations. Normalised gain is defined as $\langle g \rangle = (M_{post} - M_{pre}) / (100 - M_{pre})$, using class means on the 0–100 scale, and classified as High ($\langle g \rangle \geq 0.7$), Moderate ($0.3 \leq \langle g \rangle < 0.7$), or Low ($\langle g \rangle < 0.3$), following Hake (1998). This ratio of class mean gains is distinct from the mean of individual normalised gains and is undefined if $M_{pre} = 100$. Descriptive gains, between-class estimates, and uncertainty should be interpreted together, without treating improvement within a class as proof of the toolkit's causal effect.

2.5 Ethical Clearance

Ethical clearance. The manuscript reports that written permission to conduct the research was obtained from the school principal before the Analysis phase and that the school issued a certificate of completion after data collection. These documents establish school-level authorisation but do not by themselves establish approval by a research ethics committee or a formal exemption. No committee name, approval number, or exemption decision is supplied in the available document; the ethics-review status therefore remains to be verified. Because the study involved minors, the account must specify the consent procedure actually followed for parents or guardians, student assent, and teacher participation, including how voluntariness and any right to decline were

communicated. Written consent and assurance that participation did not affect grades should be reported only if supported by the study records.

The manuscript reports that individual responses were anonymised and presented in aggregate form. The comparison involved instruction on the same curricular content and allocated instructional time, so the control class continued to receive standard instruction. Any classroom photograph retained for publication should be checked for identifiable faces and other identifying details. Where permission for publication is documented, its scope should be stated; otherwise, the photograph should be omitted or replaced with a non-identifying illustration. Showing students from behind or at a distance does not itself establish consent for publication.

3. Results

3.1 Analysis Phase: Needs, Curriculum, and Platform Constraints

Interviews with the mathematics teacher and classroom observation conducted at the start of the 2022/2023 academic year identified two converging problems. First, smartphone use during lessons was widespread but overwhelmingly recreational rather than instructional, which fragmented students' attention during conventional explanation. Second, the printed textbook's explanation of Linear Programming was judged too dense and linguistically demanding for Grade XI readers, and teacher-produced summary handouts, while helpful, were not sufficient to lift daily-test scores above the 78-point mastery threshold for the majority of students. These findings converged on a single design implication: instruction needed to redirect students' existing device use toward structured, teacher-controlled learning activity rather than compete with it, which supported the choice of a blended rather than a purely face-to-face redesign.

A curriculum analysis mapped the mandated basic competencies for Linear Programming onto achievement indicators and, in parallel, onto the four creative-thinking indicators of fluency, flexibility, originality, and elaboration. This mapping confirmed the assumption motivating the study: Linear Programming's contextual optimisation problems can be modelled with more than one variable definition, solved through more than one constraint-handling strategy, and interpreted through more than one real-world framing, giving the topic strong intrinsic potential for creative-thinking instruction (Sugiarto et al., [2021](#); Rahman et al., [2021](#)). A needs analysis of the toolkit itself indicated that three components were required: a teaching module detailing the blended lesson scenario across face-to-face and Microsoft Teams sessions, a student worksheet built around open-ended contextual problems rather than purely procedural exercises, and a validated creative-thinking assessment aligned to the same four indicators.

A platform analysis of Microsoft Teams, conducted against the same design brief, confirmed that its assignment, file-repository, and synchronous-meeting features could jointly support the intended blended scenario, consistent with prior reports of the platform's suitability for structured mathematics instruction (Situmorang, [2020](#); Irmayanti et al., [2023](#)). Two intact Grade XI science classes were identified as suitable for the subsequent Evaluation phase on the teacher's recommendation: XI MIPA 2, selected

as the experimental class in which the toolkit would be implemented, and XI MIPA 3, selected as the control class that would continue with conventional face-to-face instruction covering the same content and time allocation. The criteria for this recommendation and the steps taken to limit allocation bias are reported in Section 2.1.

3.2 Design Phase: From Analysis Findings to Design Principles

The Analysis phase findings were translated into four working design principles, summarised in Table 1 together with the specific Analysis finding that motivated each one.

Table 1

Design Principles Derived from the Analysis Phase

Design principle	Analysis finding it addresses	Realisation in the toolkit
Redirect, rather than compete with, device use	Recreational smartphone use fragmented attention	Every face-to-face session paired with a defined Microsoft Teams component
Prioritise open-ended, multi-solution tasks	Students defaulted to a single memorised procedure	Worksheet items requiring students' own variable definitions and constraint models
Simplify instructional language and layout	Textbook explanation judged too dense for readers	Shorter explanatory text, worked contextual examples, structured layout
Keep workload realistic within existing time	Risk that ambitious redesign would be impractical	Five-session sequence fitted to the school's existing weekly allocation

The first principle operationalised the Analysis-phase finding on device use most directly: rather than restricting smartphone access, the design routed it through Microsoft Teams so that time previously spent on off-task activity became pre-reading, assignment, or discussion time. The second principle responded to the curriculum-mapping finding that Linear Programming's constrained-optimisation structure supports multiple valid solution paths (Sugiarto et al., 2021; Rahman et al., 2021), translating that structural potential into a concrete task-design rule rather than leaving it implicit. Together, these four principles directly shaped the subsequent Development phase and were used later, in the Evaluation phase, as the reference points against which effectiveness evidence was interpreted.

3.3 Development Phase: Expert Validation and Iterative Revision

The teaching module, student worksheet, instructional material, and creative-thinking test were produced and then submitted to expert review. Two subject-matter lecturers validated content, one media lecturer validated the Microsoft-Teams-based

media design, and one language teacher validated linguistic quality and readability for a senior-high audience. Table 2 summarises the content-expert validation.

Table 2

Content-Expert Validation Results

Aspect	Score	Percentage	Category
Content feasibility	51/56	91%	Very Valid
Presentation of material	42/48	88%	Very Valid
Alignment with creative-thinking indicators	27/48	56%	Not Valid
Design and instructional model	30/32	94%	Very Valid
Overall	150/184	82%	Valid

The overall content-validity score of 82% (150 of a possible 184 points pooled across the two content experts, i.e., 81.52%; see Section 2.4 for the aggregation rule) met the threshold of the Valid category (76%). However, the sub-aspect scoring lowest by a wide margin — alignment with the four creative-thinking indicators, at 56% (“Not Valid”) — was also the sub-aspect most directly tied to the study’s dependent variable. Because this study treats each Development-phase evaluation as a genuine redesign trigger rather than a pass/fail checkpoint, this finding was carried forward explicitly rather than averaged away inside the overall 82% figure.

Table 3 details the specific revisions made in direct response to the validators’ comments on this sub-aspect. Each revision targeted one of the four creative-thinking indicators individually, rather than treating “creative thinking” as a single undifferentiated design target.

Table 3

Revisions Made in Response to the Creative-Thinking-Alignment Finding

Validator finding	Revision made	Indicator targeted
Worksheet items were closed-ended and left no room for flexibility or originality	Added at least one open-ended item per worksheet (e.g., “create your own real-world problem that can be modelled with Linear Programming and solve it your own way”)	Flexibility, originality
No item required more than one solution method	Added explicit instructions requiring two solution methods (corner-point and iso-profit line) with a required efficiency comparison	Flexibility

Validator finding	Revision made	Indicator targeted
Too few elaboration items; none required deep justification	Added a closing elaboration section to every worksheet (e.g., “what happens if one constraint is removed?”; “why is the optimum always at a corner point?”)	Elaboration
Links between learning objectives were not explicit	Added a concept-flow diagram at the start of the teaching module linking prerequisite concepts to Linear Programming	Fluency (conceptual linking)
Enrichment items remained procedural rather than higher-order	Replaced procedural enrichment items with two higher-order items (analysis and evaluation level) bridging inequality systems to Linear Programming	Originality, elaboration

Table 4 summarises the media- and language-expert validation, both of which were judged Valid to Very Valid without requiring the same magnitude of revision as the content instrument.

Table 4

Media- and Language-Expert Validation Results

Validator	Aspect	Percentage	Category
Media expert	Visual display of Microsoft Teams components	95%	Very Valid
Media expert	Ease of use	92%	Very Valid
Media expert	Digital content quality	85%	Valid
Media expert	Technical function and security	88%	Valid
Media expert	Overall	90%	Very Valid
Language expert	Overall (5 sub-aspects; readability, diction, spelling, register, formality)	90%	Very Valid

Following expert revision, the limited try-out ($n = 15$) checked the readability and usability of the revised worksheet and platform workflow at the item level. Student response scores from this try-out, summarised in Table 5, were high overall (88.04%, “Very Good”), but again singled out the same underlying concern: the aspect covering LKPD (worksheet) quality and learning activities, most directly related to creative-thinking stimulation, scored the lowest of the five aspects (86.0%), even though it remained within the “Very Good” category.

Table 5

Student Response Scores from the Limited Try-Out (n = 15)

Aspect	Score	Max.	Percentage	Category
Ease of use and readability	270	300	90.0%	Very Good
Quality of material presentation	270	300	90.0%	Very Good
Worksheet (LKPD) and activity quality	258	300	86.0%	Very Good
Use of Microsoft Teams as a medium	261	300	87.0%	Very Good
Clarity of test items and instruments	156	180	86.7%	Very Good
Overall	1,215	1,380	88.04%	Very Good

That the worksheet-related aspect was also the lowest-ranked aspect in the try-out (Table 5) should be interpreted cautiously. The two instruments address different constructs: the expert instrument judged whether items were aligned with the four creative-thinking indicators, whereas the student questionnaire asked about the clarity and appeal of the worksheet activities, and 86% lies well within the Very Good category. The try-out result therefore does not confirm the construct problem identified by the experts; at most it is compatible with it. The two findings were linked more directly by students' written comments, which prompted the wording adjustments described next. Two specific try-out items, concerning creative-thinking stimulation and elaboration questions, were flagged for further minor wording adjustments before the toolkit proceeded to the Implementation phase. This two-step revision pattern — expert-level content revision followed by user-level usability revision — illustrates the kind of nested iteration that DBR studies typically report (Tinoca et al., 2022), and it meant that the Implementation phase proper could focus on instructional enactment rather than on further redesign of the materials themselves.

The creative-thinking test itself, administered as pretest and posttest to both classes, consisted of open-ended contextual items directly aligned to the four indicators used throughout this study: a fluency item asking students to propose as many distinct valid mathematical models as possible for a given optimisation scenario; a flexibility item solvable through more than one legitimate strategy (for example, graphical inspection versus corner-point evaluation); an originality item rewarding unconventional but mathematically valid reinterpretations of the given context; and an elaboration item requiring a fully justified, step-by-step optimal solution rather than a bare final answer. This item structure operationalises, at the assessment level, the same four-indicator framework that the Design phase used to redesign the worksheet, which is one reason the test could meaningfully detect the treatment's effect on creative thinking specifically rather than on general achievement alone.

3.4 Implementation Phase: Practicality in Authentic Classroom Use

The revised toolkit was implemented with the experimental class (XI MIPA 2) across five blended sessions covering the full Linear Programming unit, combining face-to-face explanation and discussion with Microsoft-Teams-based asynchronous assignments and, where scheduled, synchronous discussion windows, while the control class (XI MIPA 3) covered the same content and time allocation through conventional face-to-face instruction. Figure 3 documents part of a face-to-face session during this implementation. Figure 3

Documentation of blended-learning implementation in the experimental class



Teachers reported that the toolkit integrated readily into the existing lesson flow, while students rated the Microsoft Teams interface, the Assignment-based task-submission workflow, and the digital instructional material as sufficiently clear and not confusing. Minor technical issues that arose during the five sessions (for example, intermittent access problems on some student devices) were resolved quickly because a technical readiness check had been built into the pre-implementation preparation, and did not materially disrupt the instructional flow. This pattern is consistent with prior evidence that Microsoft Teams and comparable platforms can sustain positive student engagement when institutions and teachers implement them with adequate preparation (Khrisat & Fakhouri, 2024; Irmayanti et al., 2023).

Observation across the five sessions also indicated some unevenness in engagement: a minority of students remained more passive during small-group discussion, particularly in the earlier sessions. The field notes suggest that participation became somewhat broader in later sessions as students grew familiar with the open-ended task format introduced by the Design-phase revisions; because participation was

not counted systematically, this is reported as a tentative impression rather than an established finding. This observational pattern is consistent with the Development phase's own finding that the worksheet's creative-thinking components, although substantially strengthened, remained the aspect with the most room for further refinement (Tables 2 and 5), and it is revisited in the Discussion as part of the study's overall interpretation of how design tensions surfaced by DBR can persist, in attenuated form, even after a documented redesign cycle.

3.5 Evaluation Phase: Effectiveness for Mathematical Creative Thinking

Prerequisite analyses confirmed that parametric comparison was appropriate: Shapiro-Wilk tests indicated that pretest, posttest, and gain scores were normally distributed in both classes (all $p > .05$), and Levene's test indicated homogeneous variances across classes for pretest, posttest, and gain scores (all $p > .05$).

Table 6

Descriptive Statistics for Pretest, Posttest, and Gain

Class	Variable	n	M	SD
Experimental	Pretest	30	63.13	8.99
Experimental	Posttest	30	75.18	10.22
Experimental	Gain	30	12.05	4.96
Control	Pretest	30	63.78	9.31
Control	Posttest	30	66.68	11.01
Control	Gain	30	2.90	4.54

The pretest means of the two classes did not differ significantly ($t(58) = -0.28$, $p = .783$; mean difference = -0.65 , 95% CI $[-5.38, 4.08]$). This result does not in itself establish that the intact classes were equivalent, and unmeasured class-level differences cannot be excluded; the posttest comparison was therefore adjusted for pretest scores. The ANCOVA showed that, after adjustment, the experimental class obtained a higher posttest mean ($M_{adj} = 75.52$) than the control class ($M_{adj} = 66.34$), a difference of 9.18 points, 95% CI $[6.70, 11.65]$, $F(1, 57) = 55.13$, $p < .001$, partial $\eta^2 = .49$. In the supplementary unadjusted analyses, the experimental class scored significantly higher than the control class at posttest ($t(58) = 3.097$, $p = .003$; difference = 8.50 , 95% CI $[3.01, 13.99]$) and showed a significantly larger gain ($t(58) = 7.454$, $p < .001$; difference = 9.15 , 95% CI $[6.69, 11.61]$). Within-class paired-samples t -tests confirmed that both classes improved significantly from pretest to posttest, but far more strongly in the experimental class ($t(29) = 13.305$) than in the control class ($t(29) = 3.497$).

Table 7

Normalised Gain and Effect Size

Class	Mean pretest	Mean posttest	$\langle g \rangle$	$\langle g \rangle$ category	Cohen's d (gain) [95% CI]
Experimental	63.13	75.18	0.327 (32.7%)	Moderate	1.93 [1.32, 2.54]
Control	63.78	66.68	0.080 (8.0%)	Low	—

Applying the normalised-gain criterion (Hake, 1998), the experimental class's gain fell in the Moderate category ($\langle g \rangle = 0.33$), whereas the control class's gain fell in the Low category ($\langle g \rangle = 0.08$). Cohen's d was 1.93 (95% CI [1.32, 2.54]) for the gain-score difference and 0.80 (95% CI [0.27, 1.33]) for the posttest difference. The gain-score d should be read with caution: it is standardised by the pooled SD of gain scores (4.75), which is much smaller than the pooled SD of posttest scores (10.62) because pretest and posttest scores were highly correlated ($r \approx .87$ to $.91$), so it is not directly comparable with posttest-based effect sizes reported in meta-analyses. The small single-site sample, the involvement of the teacher-researcher, and the width of the confidence intervals also suggest that the true effect may be smaller than the point estimates. Even so, the unadjusted and pretest-adjusted results point in the same direction, indicating that the advantage associated with the toolkit was statistically reliable and practically meaningful in this classroom.

Read against the four design principles summarised in Table 1, this pattern of results is consistent with, though it cannot definitively isolate, the contribution of each principle: the large gap between the experimental and control gains is compatible with the redirection of device use and the open-ended task redesign both being active ingredients, while the moderate (rather than high) normalised-gain category for the experimental class is consistent with the Development- and Implementation-phase observation that the toolkit's creative-thinking components, although substantially strengthened, still had room for further refinement.

4. Discussion

The headline finding — a significantly larger posttest mean, a substantially larger normalised gain, and a large effect size for the experimental class relative to the control class — is consistent with the broader international evidence that blended learning tends to outperform conventional instruction when it is deliberately designed rather than adopted as a generic label. Recent meta-analytic syntheses report positive, if heterogeneous, blended-learning effects on performance and achievement across countries and educational levels (Cao, [2023](#)), with a mathematics-specific meta-analysis reporting a broadly similar pattern of moderate-to-large positive effects on math achievement (Putri et al., [2025](#)), and an emerging body of work suggesting that

technology-augmented blended designs can produce even larger gains when the technological component is used purposefully rather than incidentally (Wu et al., 2025). The present study's gain-score effect size (Cohen's $d = 1.93$) is considerably larger than the effects typically reported in this literature. Part of this difference is likely methodological: effect sizes computed on gain scores are inflated relative to posttest-based effect sizes when pretest–posttest correlations are high, and the posttest-based estimate in this study ($d = 0.80$) is closer to published ranges. The small single-site sample, teacher-researcher involvement, and the close alignment between the test and the redesigned tasks may also have inflated the estimate, which should therefore be treated as provisional until replicated.

At the same time, the magnitude of this study's effect should be read cautiously against the well-documented finding that blended learning's benefits are conditional rather than automatic. Systematic reviews of the online component of blended learning show that positive outcomes depend heavily on deliberate instructional design, scaffolding, and integration between the online and face-to-face components, rather than emerging simply from technology adoption (Rasheed et al., 2020). The bibliometric growth of the blended/hybrid learning literature documented by Bozkurt (2022) likewise signals a maturing field in which “does blended learning work” is increasingly being replaced by “under what design conditions does it work,” a framing this study's DBR structure was explicitly built to answer. Seen in that light, the Analysis-to-Design translation reported in Section 3.2 — pairing every face-to-face session with a defined Microsoft Teams component rather than layering technology on top of unchanged pedagogy — is arguably as important to the effectiveness result as the platform itself, a pattern consistent with math-specific evidence that blended designs raise both achievement and engagement chiefly when the technology is purposefully aligned with pedagogical goals (Alamri, 2021; Attard & Holmes, 2020), and with the original theoretical case for blended learning as a genuine redesign of the learning environment rather than a delivery-channel substitution (Garrison & Kanuka, 2004).

The Microsoft-Teams-specific literature situates this result further. Evidence from Jordanian universities shows that Microsoft Teams, alongside Moodle and Zoom, can meaningfully shape student engagement and achievement, but that its effect is moderated by how strategically institutions implement it (Khrisat & Fakhouri, 2024) — a pattern that mirrors this study's Implementation-phase finding that a pre-implementation technical readiness check, rather than the platform's features alone, appears to have protected instructional continuity across the five blended sessions.

The study's most theoretically interesting finding, however, is not the aggregate effectiveness result but the 56% “Not Valid” score on the creative-thinking-alignment sub-aspect uncovered during the Development phase (Table 2), and its echo in the lowest-scoring aspect of the limited try-out (Table 5). This finding matters because it would have been invisible to a single-pass development pipeline that stops at an aggregate validity percentage; it became visible only because each cycle's evaluation was treated, in DBR fashion, as a redesign trigger rather than a pass/fail gate (The Design-Based Research

Collective, 2003; Tinoca et al., 2022). That the toolkit could simultaneously be judged “Valid” overall (82%) while its most construct-relevant sub-aspect was judged “Not Valid” (56%) is a caution against relying on a single aggregate validity figure in mathematics instructional-design research, and it echoes a broader pattern in the creative-thinking literature: creative thinking is not a by-product of otherwise sound instructional materials but a specific design target that must be engineered into tasks through explicit multiple-solution, problem-posing, or open-ended structures (Leikin & Pitta-Pantazi, 2013; Wilkie, 2024; Schindler & Lilienthal, 2020). The revisions documented in Table 3 — items explicitly demanding multiple constraint formulations, alternative solution strategies, and unconventional but valid interpretations — are a direct, traceable application of this literature to the Linear Programming context, and may offer one plausible explanation for the effectiveness result, although the present design cannot isolate the contribution of the revisions from that of the teacher's facilitation or the Microsoft Teams platform, because these were evaluated together as one package.

This finding also resonates with recent evidence that structured creative-problem-solving interventions can raise students' fluency, flexibility, and originality scores when tasks are deliberately engineered for those indicators (Ersoy & Özcan Kara, 2026), and it is broadly consistent with Indonesian evidence that students' difficulty with creative, non-routine linear-inequality and Linear-Programming tasks often reflects the scarcity of such tasks in standard instructional materials rather than a fixed ceiling on students' capability (Sari & Afriansyah, 2022; Wijaya et al., 2022). Read together, these strands suggest that the toolkit's design-tension-and-redesign narrative reported here may not be an idiosyncratic feature of this particular classroom but could reflect a more general pattern in mathematics instructional design, a possibility that requires testing in other schools: creative-thinking alignment is the aspect most likely to be under-specified by default, and most in need of the kind of iterative scrutiny that DBR, by construction, provides (Anderson & Shattuck, 2012; Zheng, 2015).

Methodologically, this study also illustrates the practical value of committing to DBR's redesign-trigger logic from the outset of a development study, rather than only aggregating validation scores into a single pass/fail figure. Treating each cycle's evaluation evidence as an explicit input into the next cycle's design, in the way recommended by DBR's foundational and subsequent methodological literature (The Design-Based Research Collective, 2003; Barab & Squire, 2004; Wang & Hannafin, 2005), did not, in this study, require data beyond the validation, try-out, and test data that a conventional validity-and-practicality study would gather, although the additional revision round required extra development time; what changed was mainly how that evidence was used and reported — as a redesign trigger rather than a terminal report — and it was this reporting choice that surfaced the creative-thinking-alignment tension documented in Tables 2, 3, and 5.

The practicality result (83.33%) also deserves brief comment because practicality and effectiveness do not always move together in instructional-design research: a toolkit can be well liked without producing measurable learning gains, or vice versa. Here the

two converged, which is consistent with the view that perceived ease of use and workflow fit are preconditions for, rather than substitutes for, learning effectiveness (Khrisat & Fakhouri, 2024). Teachers' reports that the toolkit integrated readily into the existing lesson flow, and students' reports that the Assignment-based submission workflow was clear, suggest that the platform-level usability groundwork completed in the Development phase freed classroom time for the cognitively demanding, open-ended tasks that the Design-phase principles had prioritised, rather than being consumed by navigating an unfamiliar interface.

For classroom practice, these findings suggest two concrete, low-cost implications for teachers and instructional designers working in similarly resourced Indonesian secondary schools. First, before adopting any blended-learning platform for mathematics instruction, worksheet and assessment items should be explicitly audited against the four creative-thinking indicators rather than against content coverage alone, since this study's own materials passed a general content-validity check while still under-serving creative thinking specifically. Second, technology adoption decisions should be paired with an explicit mapping of platform features onto pedagogical functions — assignment, resource repository, synchronous discussion — rather than introduced as an undifferentiated “digital version” of existing materials, echoing the theoretical distinction between thoughtful integration and mere technology addition that underlies the blended-learning literature more broadly (Garrison & Kanuka, 2004; Bozkurt, 2022). For teacher-education programmes and graduate mathematics-education supervisors, the study additionally illustrates that reporting development research through an explicit redesign-trigger lens, rather than as a linear single-pass pipeline, was feasible within the data that a conventional development study already collects and, in this case, added interpretive value, a point with direct relevance for how development theses in Indonesian graduate mathematics-education programmes might be supervised and written.

Several limitations qualify these conclusions. The Evaluation phase used two intact classes allocated non-randomly on the teacher's recommendation. Although the pretest difference was small and the main comparison was adjusted for pretest scores, ANCOVA cannot remove the influence of unmeasured class-level differences (e.g., class climate, motivation, or timetable effects), so the results should be read as associations rather than as established causal effects. The sample was drawn from a single school over a single five-session unit, which limits generalisation to other schools, teachers, and mathematics topics. Finally, because the same teacher-researcher was involved in both design and implementation, some motivational bias favouring the experimental condition cannot be fully ruled out, a limitation shared with much single-site instructional-design research and one that future DBR cycles conducted by independent teachers could help address. In addition, the comparison evaluated a package of open-ended tasks, teacher facilitation, and the Microsoft Teams environment, and therefore cannot isolate the contribution of the platform or of any single design principle.

5. Conclusion

This study used Design-Based Research, from the outset, to design, validate, implement, and evaluate a Microsoft-Teams-based blended learning toolkit intended to strengthen Grade XI students' mathematical creative thinking in Linear Programming. Organised as one DBR macro-cycle with four phases and two embedded redesign micro-cycles, the study found that a needs- and curriculum-driven design produced a toolkit judged valid by content, media, and language experts (82–90%), practical in authentic classroom use (83.33%), and associated with greater improvement in students' mathematical creative thinking than conventional instruction: the experimental class had a higher pretest-adjusted posttest mean (75.52 vs. 66.34), a moderate rather than low normalised gain ($\langle g \rangle = 0.33$ vs. 0.08), and a large effect size. Because this evidence comes from two non-randomly allocated intact classes in a single school, it indicates an association rather than an established causal effect, and it cannot isolate the contribution of Microsoft Teams itself. Three provisional contributions follow. Theoretically, the study adds a mathematics-education case in which a specific instructional-design tension — the under-alignment of otherwise valid instructional materials with creative-thinking indicators — was surfaced, documented, and addressed through an explicit redesign micro-cycle rather than left implicit in an aggregate validity score. Methodologically, the study illustrates that adopting DBR's redesign-trigger logic from the start of an Indonesian graduate mathematics-education development study was feasible and, in this case, required no data beyond what a conventional validity-and-practicality study would gather, while surfacing a design tension that aggregate-score reporting would have left invisible. Practically, the study offers teachers and instructional designers provisional design principles for Microsoft-Teams-based blended mathematics instruction: pairing every face-to-face session with a defined online component, and auditing worksheet items explicitly against fluency, flexibility, originality, and elaboration indicators rather than against content coverage alone. These principles require replication in other schools, with other teachers and mathematics topics, and with randomised or better-matched comparison groups, ideally with implementation led by teachers independent of the research team, before they can be regarded as transferable.

Statements and Declarations

Funding:

This research was conducted as part of the first author's master's thesis and was self-funded. No external funding was received.

Competing Interests:

The authors declare that they have no competing interests.

Ethics Approval:

Written permission to conduct the study was obtained from the principal of SMA Kartini, Batam.

Informed Consent:

Individual responses were anonymised and reported in aggregate form.

Data Availability:

Anonymised data supporting the findings of this study are available from the corresponding author upon reasonable request, subject to participant confidentiality and institutional requirements.

Author Contributions:

Erza Candra: Conceptualisation, Methodology, Investigation, Data curation, Formal analysis, Writing – original draft.

Yumiati Yumiati: Methodology, Supervision, Writing – review & editing.

Endang Wahyuningrum: Validation, Supervision, Writing – review & editing.

Declaration of Generative AI Use:

During manuscript preparation, ChatGPT (OpenAI) was used to assist with language editing, text organisation, reference formatting, and drafting sections of the manuscript. The authors are responsible for reviewing and verifying all AI-assisted content and for the final publication.

Acknowledgements:

The authors thank the principal, mathematics teacher, and students of SMA Kartini, Batam, for their participation and cooperation, and the expert validators for their constructive feedback.

6. References

- Aiken, L. R. (1980). Content validity and reliability of single items or questionnaires. *Educational and Psychological Measurement*, 40(4), 955–959. <https://doi.org/10.1177/001316448004000419>
- Aiken, L. R. (1985). Three coefficients for analyzing the reliability and validity of ratings. *Educational and Psychological Measurement*, 45(1), 131–142. <https://doi.org/10.1177/0013164485451012>
- Alamri, M. M. (2021). Using blended project-based learning for students' behavioral intention to use and academic achievement in higher education. *Education Sciences*, 11(5), Article 207. <https://doi.org/10.3390/educsci11050207>
- Anderson, T., & Shattuck, J. (2012). Design-based research: A decade of progress in education research? *Educational Researcher*, 41(1), 16–25. <https://doi.org/10.3102/0013189X11428813>
- Attard, C., & Holmes, K. (2020). "It gives you that sense of hope": An exploration of technology use to mediate student engagement with mathematics. *Heliyon*, 6(1), Article e02945. <https://doi.org/10.1016/j.heliyon.2019.e02945>
- Barab, S., & Squire, K. (2004). Design-based research: Putting a stake in the ground. *Journal of the Learning Sciences*, 13(1), 1–14. https://doi.org/10.1207/s15327809jls1301_1
- Bozkurt, A. (2022). A retro perspective on blended/hybrid learning: Systematic review, mapping and visualization of the scholarly landscape. *Journal of Interactive Media in Education*, 2022(1), Article 2. <https://doi.org/10.5334/jime.751>
- Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative Research in Psychology*, 3(2), 77–101. <https://doi.org/10.1191/1478088706qp063oa>

- Cao, W. (2023). A meta-analysis of effects of blended learning on performance, attitude, achievement, and engagement across different countries. *Frontiers in Psychology*, 14, Article 1212056. <https://doi.org/10.3389/fpsyg.2023.1212056>
- Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Lawrence Erlbaum Associates.
- Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, 16(3), 297–334. <https://doi.org/10.1007/BF02310555>
- Dimitrov, D. M., & Rumrill, P. D., Jr. (2003). Pretest-posttest designs and measurement of change. *Work*, 20(2), 159–165.
- Ersoy, E., & Özcan Kara, D. (2026). The effect of the creative problem-solving approach on creative thinking skills. *Frontiers in Psychology*, 17, Article 1734855. <https://doi.org/10.3389/fpsyg.2026.1734855>
- Garrison, D. R., & Kanuka, H. (2004). Blended learning: Uncovering its transformative potential in higher education. *The Internet and Higher Education*, 7(2), 95–105. <https://doi.org/10.1016/j.iheduc.2004.02.001>
- Hake, R. R. (1998). Interactive-engagement versus traditional methods: A six-thousand-student survey of mechanics test data for introductory physics courses. *American Journal of Physics*, 66(1), 64–74. <https://doi.org/10.1119/1.18809>
- Hedges, L. V., & Olkin, I. (1985). *Statistical methods for meta-analysis*. Academic Press.
- Irmayanti, I., Jamaluddin, Musdalifah, A., & Sudirman, P. (2023). Efektivitas penerapan media pembelajaran Microsoft Teams terhadap minat dan hasil belajar mahasiswa. *JTMT: Journal Tadris Matematika*, 4(1), 49–55. <https://doi.org/10.47435/jtmt.v4i1.1497>
- Khrisat, Z., & Fakhouri, H. N. (2024). Impact of e-learning tools (Moodle, Microsoft Teams, Zoom) on student engagement and achievement at Jordan universities. *International Journal of Interactive Mobile Technologies*, 18(18), 125–145. <https://doi.org/10.3991/ijim.v18i18.49895>
- Leikin, R., & Pitta-Pantazi, D. (2013). Creativity and mathematics education: The state of the art. *ZDM Mathematics Education*, 45(2), 159–166. <https://doi.org/10.1007/s11858-012-0459-1>
- Levene, H. (1960). Robust tests for equality of variances. In I. Olkin (Ed.), *Contributions to probability and statistics* (pp. 278–292). Stanford University Press.
- OECD. (2023). *PISA 2022 results (Volume I): The state of learning and equity in education*. OECD Publishing. <https://doi.org/10.1787/53f23881-en>
- OECD. (2024). *PISA 2022 results (Volume III): Creative minds, creative schools*. OECD Publishing. <https://doi.org/10.1787/765ee8c2-en>
- Putri, A. D., Juandi, D., Turmudi, & Suparman. (2025). Blended learning and math achievement: A meta-analytic review highlighting the effectiveness and heterogeneity. *Electronic Journal of e-Learning*, 23(1), 113–128. <https://doi.org/10.34190/ejel.23.1.3781>
- Rahman, O., Usman, & Johar, R. (2021). Improving high school students' critical thinking ability in linear programming through problem based learning assisted by

- GeoGebra. *Journal of Physics: Conference Series*, 1882(1), Article 012070. <https://doi.org/10.1088/1742-6596/1882/1/012070>
- Rasheed, R. A., Kamsin, A., & Abdullah, N. A. (2020). Challenges in the online component of blended learning: A systematic review. *Computers & Education*, 144, Article 103701. <https://doi.org/10.1016/j.compedu.2019.103701>
- Sari, R. F., & Afriansyah, E. A. (2022). Kemampuan berpikir kreatif matematis dan belief siswa pada materi persamaan dan pertidaksamaan linear. *Plusminus: Jurnal Pendidikan Matematika*, 2(2), 275–288. <https://doi.org/10.31980/plusminus.v2i2.1755>
- Schindler, M., & Lilienthal, A. J. (2020). Students' creative process in mathematics: Insights from eye-tracking-stimulated recall interviews on students' work on multiple solution tasks. *International Journal of Science and Mathematics Education*, 18(8), 1565–1586. <https://doi.org/10.1007/s10763-019-10033-0>
- Shapiro, S. S., & Wilk, M. B. (1965). An analysis of variance test for normality (complete samples). *Biometrika*, 52(3/4), 591–611. <https://doi.org/10.2307/2333709>
- Situmorang, A. S. (2020). Microsoft Teams for Education sebagai media pembelajaran interaktif meningkatkan minat belajar. *Sepren*, 2(1), 30–37. <https://doi.org/10.36655/sepren.v2i1.351>
- Spatioti, A. G., Kazanidis, I., & Pange, J. (2022). A comparative study of the ADDIE instructional design model in distance education. *Information*, 13(9), Article 402. <https://doi.org/10.3390/info13090402>
- Sugiarto, I. J., Usodo, B., & Saputro, D. R. S. (2021). High school students' mathematic literacy performance in solving linear programming problem. *Journal of Physics: Conference Series*, 1776(1), Article 012014. <https://doi.org/10.1088/1742-6596/1776/1/012014>
- The Design-Based Research Collective. (2003). Design-based research: An emerging paradigm for educational inquiry. *Educational Researcher*, 32(1), 5–8. <https://doi.org/10.3102/0013189X032001005>
- Tinoca, L., Piedade, J., Santos, S., Pedro, A., & Gomes, S. (2022). Design-based research in the educational field: A systematic literature review. *Education Sciences*, 12(6), Article 410. <https://doi.org/10.3390/educsci12060410>
- Van Breukelen, G. J. P. (2006). ANCOVA versus change from baseline had more power in randomized studies and more bias in nonrandomized studies. *Journal of Clinical Epidemiology*, 59(9), 920–925. <https://doi.org/10.1016/j.jclinepi.2006.02.007>
- Wang, F., & Hannafin, M. J. (2005). Design-based research and technology-enhanced learning environments. *Educational Technology Research and Development*, 53(4), 5–23. <https://doi.org/10.1007/BF02504682>
- Wijaya, A. J., Pujiastuti, H., & Hendrayana, A. (2022). Tingkat kemampuan berpikir kreatif siswa dalam menyelesaikan soal open ended. *JIPM (Jurnal Ilmiah Pendidikan Matematika)*, 11(1), 108–122. <https://doi.org/10.25273/jipm.v11i1.10866>

- Wilkie, K. J. (2024). Creative thinking for learning algebra: Year 10 students' problem solving and problem posing with quadratic figural patterns. *Thinking Skills and Creativity*, 52, Article 101550. <https://doi.org/10.1016/j.tsc.2024.101550>
- Wu, J., Tlili, A., Salha, S., Mizza, D., Saqr, M., López-Pernas, S., & Huang, R. (2025). Unlocking the potential of artificial intelligence in improving learning achievement in blended learning: A meta-analysis. *Frontiers in Psychology*, 16, Article 1691414. <https://doi.org/10.3389/fpsyg.2025.1691414>
- Zheng, L. (2015). A systematic literature review of design-based research from 2004 to 2013. *Journal of Computing in Education*, 2(4), 399–420. <https://doi.org/10.1007/s40692-015-0036-z>