

Research Article

The weak link between classroom gamification and scientific literacy: A multi-method correlational study of senior high school students in a green chemistry context

Dewi Satria Ahmar¹, Muhammad Fath Azzajjad², Tri Santoso³ and Muhammad Ismail⁶

¹Department of Chemistry Education, Faculty of Teacher Training and Education, Tadulako University, Indonesia (ORCID: 0000-0003-2537-3674)

²Department of Chemistry Education, Faculty of Teacher Training and Education, Universitas Sembilanbelas November Kolaka, Indonesia (ORCID: 0000-0002-9131-3437)

³Department of Chemistry Education, Faculty of Teacher Training and Education, Tadulako University, Indonesia (ORCID: 0000-0003-0959-3626)

⁴Department of Sports Coaching Education, Faculty of Teacher Training and Education, Tadulako University, Indonesia (ORCID: 0009-0000-5525-8031)

Gamification, the use of game design elements such as points, badges, and competition in classroom instruction, is widely promoted as a strategy to raise student motivation and, by extension, learning outcomes, yet empirical evidence for its downstream effect on measured achievement remains mixed. This study examined the relationship between classroom gamification and scientific literacy among senior high school students across eight schools in Palu, Central Sulawesi, Indonesia, using a durian-waste, green-chemistry context as the literacy assessment theme. A gamification questionnaire and a three-tier diagnostic scientific literacy test were administered to the same students, and four complementary analytical strategies, exploratory factor analysis, cluster analysis, correlation and network analysis, and structural equation modeling, were applied to examine whether gamification experience predicts scientific literacy achievement and misconception profile, and whether this relationship depends on the school a student attends. The gamification instrument proved to be empirically closer to a single general engagement construct than to the four-dimension structure assumed at the outset, and one of the four hypothesized subscales did not meet conventional reliability standards. Across all four analytical strategies, gamification experience showed no reliable association with scientific literacy achievement, whether examined through total score, engagement-based student groupings, item-level network structure, or a latent structural model; a school-level moderation test likewise found no evidence that this null relationship differed across schools. Gamification exposure, engagement, and enjoyment, as reported by students, were therefore not, by themselves, sufficient predictors of scientific literacy achievement in this setting, and any relationship that does exist more plausibly operates through how consistently gamified instruction is implemented in the classroom than through student-reported experience alone. Implications for context-based, green-chemistry science instruction, and for the measurement of both classroom gamification and scientific literacy in similar settings, are discussed.

Keywords: Gamification; Green chemistry education; Scientific literacy; Structural equation modeling; Three-tier diagnostic test

Article History: Submitted 9 August 2026; Revised 26 September 2026; Accepted 29 September 2026; Published online 8 October 2026

1. Introduction

Classroom gamification, defined as the use of game design elements such as points, badges, levels, and competition in a non-game instructional setting (Deterding et al., 2011), has become one of the

Address of Corresponding Author

Dewi Satria Ahmar, Department of Chemistry Education, Faculty of Teacher Training and Education, Tadulako University, Indonesia.

✉ dewisatriaahmar@gmail.com

How to cite: Ahmar, D. S., Azzajjad, M. F., Santoso, T., & Ismail, M. (2026). The weak link between classroom gamification and scientific literacy: A multi-method correlational study of senior high school students in a green chemistry context. *Journal of Pedagogical Research*, 10(4), 133-156. <https://doi.org/10.33902/JPR.202648842>

most widely adopted pedagogical strategies of the last decade (Kapp, 2012; Seaborn & Fels, 2015). Systematic reviews report that gamified instruction reliably increases short-term engagement and self-reported motivation across educational levels (Ratinho & Martins, 2023), and gamification-specific reviews in science education report similar patterns of increased interest and participation (Kalogiannakis et al., 2021). Individual intervention studies extend this pattern into laboratory and applied science settings: gamifying physics laboratory sessions has been shown to raise both motivation and skill acquisition relative to non-gamified sessions (Okariz et al., 2023), and a gamified platform designed for scientific writing produced measurable gains in writing performance among engineering students under a quasi-experimental design (Nunez-Pacheco et al., 2023).

The theoretical rationale for expecting gamification to improve learning outcomes, not merely engagement, rests substantially on self-determination theory [SDT]. SDT holds that intrinsically motivated learning follows from the satisfaction of three psychological needs: autonomy, competence, and relatedness (Deci & Ryan, 1985; Ryan & Deci, 2000). Game elements are assumed to support these needs by offering choice, immediate feedback on competence, and social comparison. However, SDT itself contains a caution that is frequently omitted from gamification advocacy: externally imposed rewards and controlling feedback structures, precisely the kind that points and leaderboards can create, may undermine rather than support intrinsic motivation (Deci et al., 1999). This tension implies that gamification's effect on deep, transferable learning outcomes such as scientific literacy is not guaranteed even when its effect on surface engagement is well documented, and reviews of the gamification literature increasingly call this distinction the field's central unresolved question (Dichev & Dicheva, 2017).

Scientific literacy, the outcome variable of interest in the present study, is defined by the OECD Programme for International Student Assessment [PISA] as the capacity to engage with science-related issues and with the ideas of science as a reflective citizen, encompassing the ability to explain phenomena scientifically, evaluate scientific inquiry, and interpret data and evidence (OECD, 2006). Indonesian students have consistently been assessed within this framework, and PISA science assessment design has continued to evolve toward broader, context-embedded item formats (OECD, 2023). Reflecting this trajectory, a recent bibliometric analysis of the science-education literature identified scientific literacy, alongside sustainability and socioscientific issues, as one of the field's most active current research fronts, reinforcing the rationale for studying it in a locally grounded, real-world context such as the one adopted here (Arifin et al., 2025). A common strategy for building scientific literacy at the classroom level is context-based learning, in which abstract chemistry content is anchored to a locally meaningful real-world problem so that students can connect classroom knowledge to lived experience (Aikenhead, 2006). Central Sulawesi is a major durian-producing region, and durian peel and seed waste, an abundant local agricultural byproduct, has documented value as a raw material for activated carbon and biochar used in water treatment (Do et al., 2025). This combination of local relevance and genuine green-chemistry application motivated the choice of a durian-waste context for the scientific literacy instrument used in this study.

A second measurement concern in this study is the diagnosis of student misconceptions, which is treated in the science education literature as distinct from simple incorrect answers. A student who guesses incorrectly because of missing knowledge is not equivalent, pedagogically, to a student who confidently and consistently applies an incorrect mental model. The three-tier diagnostic test, which layers a content question, a reasoning-tier question, and a certainty-of-response rating, was designed specifically to separate lack of knowledge from genuine misconception (Hasan et al., 1999; Treagust, 1988). The approach has been validated across a range of chemistry topics, including acid-base chemistry (Caleon & Subramaniam, 2010), chemical equilibrium (Bilgin et al., 2003), states of matter (Kirbulut & Geban, 2014), titration phenomena (Sheppard, 2006), and chemical bonding (Chandrasegaran et al., 2007; Vrabec & Proksa, 2016), and

remains one of the most widely used instruments for capturing the depth, rather than merely the presence, of student misunderstanding (Taber, 2002).

Despite the breadth of research on gamification and on scientific literacy separately, few studies have examined whether the well-documented motivational effect of gamification actually propagates into measured scientific literacy achievement and misconception profiles when both variables are collected from the same students in a single classroom setting, rather than compared across a treatment and a control group. The present study addresses this gap directly. Using survey and diagnostic-test data collected from 343 senior high school students across eight schools in Palu, Central Sulawesi, this study asks: (1) what is the empirical factor structure and typology of student-reported gamification experience in this sample; (2) does gamification experience, at the individual or the dimensional level, predict scientific literacy achievement and misconception profile; and (3) does the strength of this relationship depend on the school in which a student is enrolled. Four complementary analytical strategies, exploratory factor analysis, cluster analysis, correlation and network analysis, and structural equation modeling (SEM), are applied to the same dataset so that any relationship, or its absence, can be corroborated across methods rather than resting on a single statistical test.

2. Literature Review and Hypotheses

The dominant empirical pattern in the gamification literature is a reliable short-term increase in engagement and motivation that is not matched by an equally reliable increase in measured achievement (Ratinho & Martins, 2023). Several explanations for this pattern have been proposed. First, gamification effects appear to depend heavily on implementation fidelity: game elements introduced inconsistently or only occasionally, as opposed to being embedded systematically into the instructional sequence, are less likely to translate motivational gains into learning gains (Kalogiannakis et al., 2021). This planning-dependence is not unique to K-12 science settings: a case study of a redesigned higher-education course found that gamification's classroom benefits depended on a deliberate, game-thinking approach to instructional design rather than the addition of game elements alone, and warned that a narrow, ad hoc view of gamification both limits its potential and risks counterproductive outcomes (Hellberg, 2023). A parallel argument has been made for inquiry-based learning models more broadly: a systematic review found that the learning benefit of an instructional model depends heavily on how faithfully its component phases are implemented, not merely on which model is nominally adopted (Fadillah et al., 2025). Second, the psychometric assumption that a gamification questionnaire measures several distinct sub-constructs, commonly utilization, engagement, motivation, and experience, is rarely tested empirically before being used to build composite scores, even though violations of this assumption directly weaken any subsequent structural or correlational analysis (Fabrigar et al., 1999). Third, most published studies that report a positive gamification effect on achievement use a quasi-experimental design with a non-gamified control group (Okariz et al., 2023; Nunez-Pacheco et al., 2023), which isolates the effect of introducing gamification; a purely correlational design conducted after the fact, in a setting where gamification exposure already varies naturally and inconsistently across teachers and classrooms, is a substantially weaker test of the same claim.

The measurement of scientific literacy through a three-tier diagnostic test embedded in a context-based, green-chemistry scenario draws on two further literatures. Context-based chemistry instruction is intended to increase both interest and conceptual understanding by grounding abstract content in a scenario the student can recognize from daily life (Aikenhead, 2006), and durian-waste valorization, as an emerging line of applied green chemistry research (Do et al., 2025), offers a scenario with clear local relevance for students in Central Sulawesi. The three-tier format itself is intended to separate three distinct student states on any given item, namely correct understanding, partial or weak misconception, and strong, confidently held misconception, rather than treating all incorrect answers as equivalent (Hasan et al., 1999). This distinction matters for the present study because gamification, even where it raises engagement, plausibly affects surface-

level participation (attempting more items, discussing answers more actively) without necessarily correcting the deeper conceptual structures that a three-tier instrument is designed to detect.

Based on this literature and consistent with the SDT-derived tension noted above, four hypotheses were formulated for empirical testing rather than assumed as findings:

H1: Student-reported gamification engagement is positively associated with scientific literacy achievement score.

H2: Student-reported gamification engagement is associated with a more favorable misconception profile (a higher proportion of no-misconception responses).

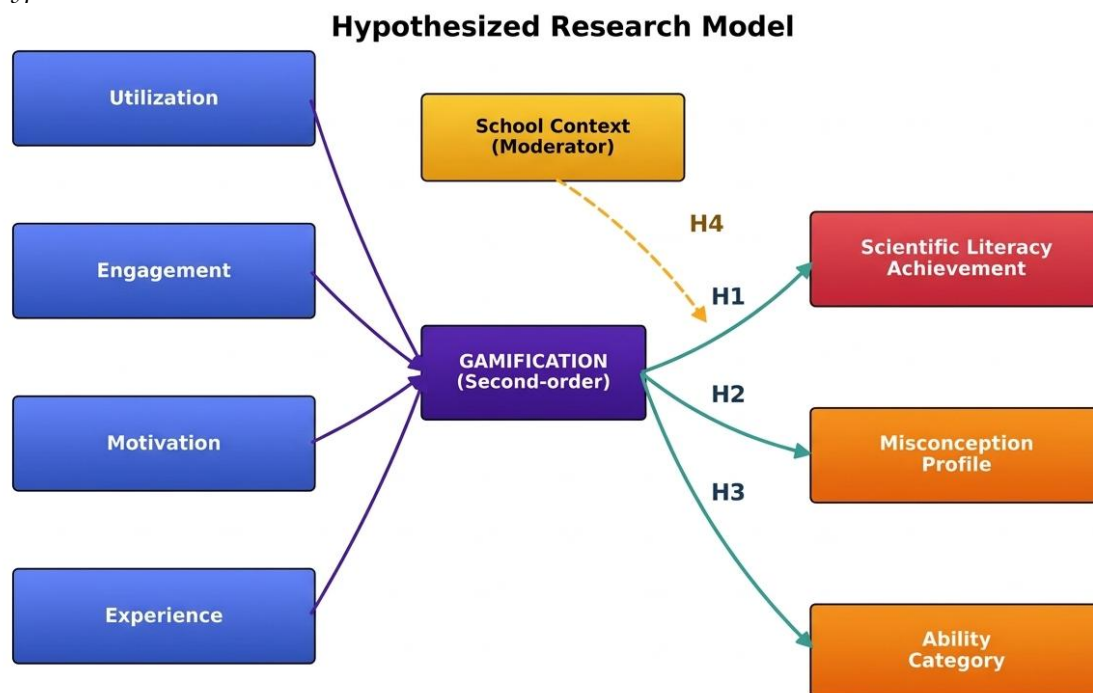
H3: Student-reported gamification engagement is associated with a higher scientific literacy ability category.

H4: The strength of the gamification-literacy relationship is moderated by school (i.e., varies across schools, plausibly reflecting differences in implementation fidelity).

The hypothesized model, including the second-order gamification construct assembled from four first-order dimensions, is presented in Figure 1. As reported in Section 4.5, the second-order structure shown in Figure 1 was not estimated in the final SEM: the exploratory factor analysis in Section 4.2 did not support treating the four dimensions as indicators of a single higher-order factor, so the four first-order factors were instead specified as direct, correlated predictors of the literacy score, without a second-order gamification factor above them. Figure 1 is retained as originally hypothesized, consistent with H1-H4 as stated, and the deviation between the hypothesized second-order model and the estimated first-order model is reported explicitly here rather than silently, in the interest of transparent reporting of a specification decision made after seeing the data.

Figure 1

Hypothesized research model



H1: Gamification predicts Scientific literacy achievement | H2: Gamification predicts Misconception profile
H3: Gamification predicts Ability category | H4: School context moderates the relationship between Gamification and Scientific Literacy Achievement

Note. The model links the four dimensions of gamification—utilization, engagement, motivation, and experience—to scientific literacy outcomes through a second-order gamification construct. H1 = gamification → scientific literacy achievement; H2 = gamification → misconception profile; H3 = gamification → ability category; H4 = school context moderates the relationship between gamification and scientific literacy achievement. The second-order gamification construct represented here was specified a priori; however, this structure was not retained in the estimated model (see Section 4.2).

3. Method

3.1. Research Design

This study used a non-experimental, cross-sectional, correlational (ex post facto) design. Data on gamification experience and on scientific literacy were collected from the same students at a single point in time for each student (schools were visited in sequence; Section 3.4), with no manipulation of instruction and no control group. This design permits tests of association and, through structural equation modeling, tests of a hypothesized directional path, but it does not permit causal inference. This limitation is stated explicitly here because it directly constrains how the results in Section 4 may be interpreted.

3.2. Participants

Participants were 343 senior high school students (matched across the two instruments from an initial pool of 344) drawn from eight public and private senior high schools in Palu, Central Sulawesi, Indonesia. Table 1 shows the distribution of the matched sample across schools.

Table 1

Distribution of the matched sample and fieldwork period by school

<i>School</i>	<i>n</i>	<i>% of sample</i>	<i>Fieldwork period (2026)</i>	<i>Assignment letter no. (date)</i>
Madani Integrated Model Public Senior High School, Palu	67	19.5	8-12 June	444 (5 June)
Public Senior High School 1, Palu	64	18.7	6-10 July	665 (2 July)
Public Senior High School 4, Palu	53	15.5	20-24 July	835 (15 July)
Public Senior High School 5, Palu	40	11.7	27-31 July	976 (27 July)
Public Senior High School 2, Palu	36	10.5	10-15 August	1185 (7 August)
Public Senior High School 3, Palu	29	8.5	13-17 July	773 (10 July)
Karuna Dipa Senior High School, Palu	30	8.7	2-5 June	380 (26 May)
Public Senior High School 9, Palu	24	7.0	3-7 August	1051 (30 July)
Total	343	100.0	2 June-15 August	8 letters

Note. Sample sizes reflect students with complete, matched records on both the gamification questionnaire and the scientific literacy test. Fieldwork periods are the assignment windows issued for each school; they also covered coordination and other project activities. Letter numbers are followed by /DST/UN28.16/KP.10.00/2026; dates are the 2026 issue dates of the letters.

Data collection followed the schools' standard research-participation procedures, and identifying information was removed prior to analysis. The fieldwork was conducted under official assignment letters issued by the university research institute for each school (see Table 1). Written parental/guardian informed consent was obtained for all participating students prior to data collection, and student assent was additionally obtained at the point of survey and test administration. The assignment letters and any consent documentation are available from the corresponding author upon request.

3.3. Instruments

3.3.1. Gamification questionnaire

Gamification experience was measured with a 17-item questionnaire on a four-point Likert scale (1 = strongly disagree to 4 = strongly agree). The items were grouped a priori into four dimensions: Utilization (3 items), Engagement (5 items), Motivation (4 items), and Experience (5 items). The instrument was developed through a combination of literature-based adaptation and contextualization for Indonesian senior high school chemistry classrooms. The 17 items and the four a priori dimensions (Utilization, Engagement, Motivation, and Experience) were adapted primarily from three sources: (1) the gamification engagement framework proposed by Hamari et al. (2014) and further refined in educational settings by Sailer et al. (2017), which informed the

Engagement and Experience dimensions; (2) the Motivational Affordances of Gamification scale and related constructs drawn from Self-Determination Theory (Deci & Ryan, 1985; Ryan & Deci, 2000) as operationalized in classroom gamification studies (e.g., Kalogiannakis et al., 2021; Ratinho & Martins, 2023), which guided the Motivation dimension; and (3) selected items from Indonesian-adapted classroom engagement and instructional-method questionnaires used in previous science-education studies in Sulawesi (with language and context adjustments for chemistry learning). The Utilization dimension (3 items) was constructed to capture the frequency and presence of game-element use by teachers, reflecting the implementation-fidelity concern highlighted in systematic reviews (Kalogiannakis et al., 2021).

All items were written in Bahasa Indonesia, reviewed for clarity and cultural appropriateness, and then subjected to expert judgment (three validators: one chemistry education specialist, one green-chemistry content expert, and one educational measurement specialist) using Aiken's V. Content validity coefficients ranged from 0.89 to 1.00 (all items rated Valid to Very Valid; see Section 3.3.3 and Table 2 for the item-level values). A pilot test with 32 senior high school students outside the main sample confirmed item clarity, acceptable item statistics (difficulty and discrimination indices within recommended ranges), and internal consistency (Cronbach's $\alpha = .81$ for the literacy instrument and $\alpha = .87$ for the full gamification scale in the pilot sample). Minor wording revisions were made after the pilot before the main administration. The questionnaire went through the same expert judgment as the literacy test (Section 3.3.3) and was pilot tested with it (Section 3.3.4).

3.3.2. Scientific literacy test

Scientific literacy was measured with a ten-item three-tier diagnostic test embedded in a durian-waste, green-chemistry reading passage. Each item combined a content-tier question, a reasoning-tier question, and a certainty-of-response rating, following the logic of Hasan et al. (1999), and student responses were coded into one of four categories per item: no misconception (correct answer, correct reasoning, high certainty), partial misconception, strong misconception, or no conception (Caleon & Subramaniam, 2010; Taber, 2002). A total literacy score (0-40, converted to a percentage) and an overall ability category (Very Low to Very High, based on the mean and standard deviation of the sample) were also derived.

The literacy test was built around a single durian-waste reading passage covering the composition of durian peel and seed, the environmental cost of untreated disposal, and documented green-chemistry valorization pathways (activated carbon and biochar production, natural dye and indicator extraction, gluten-free flour, and bioplastic/biofoam applications), with each of the ten items mapped in a test blueprint (kisi-kisi) to a specific instructional objective and scientific-literacy competency (explaining phenomena scientifically, interpreting data and evidence, or evaluating scientific inquiry, following the PISA framework cited above) and an explicit answer key for each of the three tiers. Each item was also assigned to one of four literacy levels (nominal, conceptual, functional, or multidimensional; Table 2).

3.3.3. Content validation by expert judgment

Content validity of both instruments was established through expert judgment before the pilot test. Three validators formed the panel: a chemistry education lecturer specializing in assessment, with at least ten years of experience developing three-tier and scientific literacy instruments (Expert 1); a lecturer in chemistry and green chemistry whose work centers on biomass waste utilization (Expert 2); and an educational evaluation and measurement specialist experienced in validating Likert-type and diagnostic instruments (Expert 3). Each expert rated every item of the literacy test on a four-point scale (1 = not relevant or unclear, 2 = less relevant or less clear, 3 = relevant or clear, 4 = very relevant or very clear). The ratings considered five aspects: alignment with the learning objectives and test blueprint, accuracy of the scientific concepts, clarity of language and wording, suitability of the three-tier format (answer, reasoning, and confidence

rating), and relevance of the green chemistry and durian context. The 17 gamification items were rated on relevance, language clarity, and fit with the four constructs.

Agreement was quantified with Aiken's V (Aiken, 1985): $V = \sum s / [n(c - 1)]$, where $s = r - l_0$, r is an expert's rating, l_0 is the lowest possible rating (1), n is the number of experts (3), and c is the number of rating categories (4). With three experts and four categories, V can take only the values 1.00, .89, .78, and so on in steps of .11, so a single rating below 4 moves an item below 1.00. Items were classed as very valid when $V \geq .92$, valid when $.75 \leq V < .92$, and in need of revision when $V < .75$. In this design the very-valid category therefore requires unanimous top ratings.

Table 2 lists the ratings and V values for the ten literacy items.

Table 2

Expert judgment of the ten literacy items (Aiken's V, three experts)

Item	Literacy level	Content focus	Expert ratings	V	Category	Follow-up
1	Nominal	Electron configuration and the periodic table	4, 4, 4	1.00	Very valid	Approved
2	Nominal	Green chemistry (definition)	4, 4, 4	1.00	Very valid	Approved
3	Conceptual	Electron configuration of Ca	4, 4, 3	.89	Valid	Minor rewording of reasoning tier
4	Conceptual	Green chemistry principle: renewable feedstocks	4, 4, 4	1.00	Very valid	Approved
5	Multidimensional	CO ₂ emissions and green chemistry	4, 4, 4	1.00	Very valid	Approved
6	Conceptual	Waste prevention principle	4, 4, 4	1.00	Very valid	Approved
7	Functional	Position of Mg in the periodic table	4, 4, 3	.89	Valid	Minor rewording of reasoning options
8	Multidimensional	Group and period of elements	4, 4, 4	1.00	Very valid	Approved
9	Conceptual	Environmentally benign solvents	4, 4, 4	1.00	Very valid	Approved
10	Multidimensional	SDGs and circular economy	4, 4, 4	1.00	Very valid	Approved

Note. Ratings are listed for experts 1, 2, and 3. $V = \sum s / [n(c - 1)]$ with $n = 3$ and $c = 4$. Very valid: $V \geq .92$; valid: $.75 \leq V < .92$; revise: $V < .75$.

Aiken's V ranged from .89 to 1.00 (mean = .98). Eight items received the top rating from all three experts ($V = 1.00$, very valid). Items 3 and 7 each received one rating of 3 ($V = .89$), which places them in the valid category, and the experts asked for minor rewording of the reasoning tier (item 3) and of the reasoning options (item 7). These edits were made, and no item was dropped. All 17 gamification items reached the very-valid criterion ($V \geq .92$) on relevance, language clarity, and fit with the four constructs, and none was removed. Content validity therefore supports using both instruments in the pilot. As Section 5.2 discusses, it does not show whether the gamification items separate into the four intended dimensions.

3.3.4. Pilot testing

Both instruments were then pilot tested to check language and instructions, item statistics, internal consistency, and administration time before the main data collection. The pilot involved 32 students in grades X and XI from Public Senior High School 1, Palu and Madani Integrated Model Public Senior High School, Palu, two weeks before the main data collection. All 32 students completed both instruments. These students were from the same grade levels (Grades X and XI) as the main sample but were deliberately selected from different classes that did not participate in the

main data collection. Students who took part in the pilot were therefore excluded from the final matched sample of 343 students reported in Table 1. The pilot was administered two weeks prior to the main data collection (approximately mid-may, 2026).

For each literacy item, the difficulty index (P) and the discrimination index (D) were computed. Difficulty was classed as medium at $.30 \leq P \leq .70$ and easy at $.71 \leq P \leq .85$. Discrimination was classed as good at $D \geq .40$ and very good at $D \geq .50$. Table 3 reports the results.

Table 3

Pilot item statistics for the ten literacy items (N = 32)

Item	P	Difficulty category	D	Discrimination category
1	.72	Easy	.48	Good
2	.66	Medium	.52	Very good
3	.59	Medium	.41	Good
4	.78	Easy	.45	Good
5	.53	Medium	.56	Very good
6	.69	Medium	.50	Very good
7	.63	Medium	.47	Good
8	.56	Medium	.53	Very good
9	.75	Easy	.44	Good
10	.61	Medium	.49	Good
Mean (range)	.65 (.53-.78)		.49 (.41-.56)	

Note. P = difficulty index; D = discrimination index. Difficulty: medium = .30-.70, easy = .71-.85. Discrimination: good $\geq .40$, very good $\geq .50$.

Difficulty ranged from .53 to .78 ($M = .65$). Seven items were medium and three (items 1, 4, and 9) were easy. Discrimination ranged from .41 to .56 ($M = .49$): items 2, 5, 6, and 8 were very good and the other six were good. All ten items met the criteria, so none was dropped or substantially revised.

Cronbach's alpha was .81 for the ten-item literacy score (0-40) and .87 for the 17-item gamification questionnaire. Both exceed the conventional .70 threshold (Nunnally & Bernstein, 1994) and fall in the high range ($\geq .80$) by the criterion used here. Pilot reliability is reported for the full scales only, not for the four gamification dimensions.

The practical results were favorable. Thirty of the 32 students (94%) found the instructions easy to understand, students understood the answer, reasoning, and confidence sequence of the three-tier format, and no ambiguous answer options were found. The pilot also indicated that the local durian context helped students understand the passage. Completion took a mean of 47 minutes (range 38 to 55), and a 50-minute session was set for the main administration. No change was made after the pilot, and both instruments were judged ready for the main data collection ($N = 343$).

3.4. Data Collection Procedure

Data were collected school by school between 2 June and 15 August 2026. Table 1 gives the fieldwork period at each school. The periods ran four to six calendar days, and no two schools were visited at the same time. The university research institute issued an assignment letter naming the research team and the school no later than the first day of each period (Table 1). The letters assigned the team to the schools; the ethics and consent position is stated in Section 3.2. Within each period, the team coordinated with school staff, explained the study, and administered the gamification questionnaire and the scientific literacy test to students. The same two instruments were used at every school. Of the 343 students, 115 were in grade X and 228 in grade XI. Grade was uniform within five schools (all grade X at Madani Integrated Model Public Senior High School; all grade XI at Public Senior High Schools 1, 2, 3, and 5) and mixed at Karuna Dipa Senior High School, Public Senior High School 4, and Public Senior High School 9.

The assignment letters also covered classroom observation, interviews, teacher questionnaires, and an instrument on green chemistry understanding, which belong to the wider project. The present study analyzes only the student gamification questionnaire and the scientific literacy test.

3.5. Analytical Strategy

Note that H1-H3 are each operationalized more than once across the analyses below: H1 is tested with the continuous total gamification score (correlation, network) and, indirectly, with cluster membership as a coarsened version of the same score; H2 and H3 use cluster membership only; and the SEM in Section 4.5 represents the gamification construct a third way, as four latent factors rather than a single score. These are treated here as complementary operationalizations of the same underlying construct, chosen to check whether the null pattern found with one representation (e.g., the continuous score) also holds under a categorical representation (cluster) and a latent-variable representation (SEM), rather than as independent tests of distinct constructs; this is stated explicitly here so the analytic strategy is not read as treating total score, cluster membership, and the four latent factors as interchangeable measures of separate things. Prior to hypothesis testing, internal consistency of the gamification questionnaire was assessed with Cronbach's alpha, both for the full scale and for each of the four a priori dimensions. Sampling adequacy for factor analysis was assessed with the Kaiser-Meyer-Olkin [KMO] statistic and Bartlett's test of sphericity, and the number of underlying factors was determined jointly by the Kaiser criterion (eigenvalue greater than 1; Kaiser, 1974) and visual inspection of the scree plot, following current best-practice recommendations for exploratory factor analysis (Costello & Osborne, 2005; Fabrigar et al., 1999). A varimax-rotated three-factor solution was then extracted for interpretation.

Student typologies were derived using K-means cluster analysis on the standardized 17-item response matrix, with the number of clusters selected using the average silhouette width criterion (Rousseeuw, 1987), consistent with standard cluster-validation practice (Everitt et al., 2011). H1 was tested with Pearson and Spearman correlations between total gamification score and literacy percentage score, supplemented by a network analysis of item-level Spearman correlations (thresholded at $|r| \geq .30$) to visualize the connectivity, or lack of connectivity, between the gamification item set and the literacy outcome (Epskamp et al., 2018), following the general logic that psychological constructs can be represented as networks of mutually reinforcing observed variables rather than only as reflective latent factors (Borsboom, 2017). H2 was tested with a chi-square test of independence between cluster membership and misconception profile category, and H3 was tested with a chi-square test between cluster membership and ability category, with effect size reported as Cramer's V (Cohen, 1988); the assumption of expected cell counts of at least five was checked in every case. H4 was tested using ordinary least squares regression with a School $\times$ Gamification-score interaction term. Finally, a structural equation model was specified with four first-order latent gamification factors (Utilization, Engagement, Motivation, Experience) predicting the observed literacy percentage score, estimated using maximum likelihood in the *semopy* package for Python, with model fit evaluated against the combined-index guidelines of Hu and Bentler (1999) and standard SEM practice (Anderson & Gerbing, 1988; Byrne, 2010; Kline, 2016). All analyses were conducted in Python 3.12 using *pandas*, *scikit-learn*, *factor_analyzer*, *semopy*, and *networkx*, with descriptive procedures cross-checked against standard treatments in Field (2013), Hair et al. (2010), and Tabachnick et al. (2019).

The structural model was also estimated with partial least squares SEM [PLS-SEM] as a composite-based check on the covariance-based result (Hair et al., 2022). The algorithm used the path weighting scheme with mode A (reflective) outer weights on standardized indicators and a convergence tolerance of 10^{-7} . Two specifications were fitted. Model M1 repeated the four first-order gamification constructs of Section 4.5. Model M2 used one gamification construct built from all 17 items, following the dominant factor in Section 4.2. In both, scientific literacy was a reflective construct with ten indicators, the item-level conception scores (0 = no conception, 1 = strong misconception, 2 = partial misconception, 3 = no misconception). Scores that also separate

confidence levels (3 versus 4 for correct answers) cannot be recovered from the categorized item data, so they were not used. The 47 incomplete-data item responses (1.4% of 3,430) were replaced by the indicator mean.

Measurement quality was judged by outer loadings, Cronbach's alpha, rhoA (Dijkstra & Henseler, 2015), composite reliability, and average variance extracted (AVE), and discriminant validity by the Fornell-Larcker criterion (Fornell & Larcker, 1981) and the heterotrait-monotrait ratio (HTMT; Henseler et al., 2015). Structural results were tested with 5,000 bootstrap resamples (percentile 95% confidence intervals and two-tailed *t*-tests with $n - 1$ degrees of freedom) and reported with R^2 , f^2 , and inner variance inflation factors. Because composites built with mode A weights are fitted toward their outcome, a permutation test (5,000 shuffles of the literacy block) supplied the distribution of R^2 and path coefficients expected without any association. Predictive relevance was assessed with PLSpredict (Shmueli et al., 2019; ten folds, ten repetitions) against a linear-model benchmark and the training-sample mean. Four robustness analyses used the total literacy score as a single indicator (S1), incomplete responses coded 0 (S2), indicators centered within school (S3), and within-school centering with a grade indicator (S4). The routine was written in Python by the authors and not run in SmartPLS. It was checked against the fixed-point condition of the algorithm and against simulated data with known parameters.

4. Results

4.1. Reliability and Descriptive Statistics

The full 17-item gamification scale showed excellent internal consistency (Cronbach's alpha = .91), close to the pilot estimate of .87 (Section 3.3.4). At the dimension level, reliability was uneven: Engagement (alpha = .76), Motivation (alpha = .75), and Experience (alpha = .79) reached acceptable levels, and Utilization fell well below the conventional .70 threshold (alpha = .44), indicating that this three-item subscale should not be treated as a reliable stand-alone measure in its current form (Nunnally & Bernstein, 1994). Table 4 summarizes these results together with the overall descriptive statistics for the literacy outcome.

Table 4

Reliability and descriptive statistics for the gamification dimensions and the literacy score

<i>Dimension</i>	<i>Items</i>	<i>Cronbach's alpha</i>	<i>M</i>	<i>SD</i>
Utilization	3	.44	2.62	0.56
Engagement	5	.76	2.76	0.55
Motivation	4	.75	2.93	0.61
Experience	5	.79	2.82	0.56
Full gamification scale	17	.91	2.79	0.50
Scientific literacy score (%)	10 items	-	48.9	19.6

On the literacy outcome, 4.1% of students fell in the Very Low ability category, 32.4% Low, 31.2% Medium, and the remainder Medium-High to Very High. On the misconception profile, 41.7% of students showed no misconception across the ten items, 31.5% showed a partial/weak misconception profile, and 12.8% showed a strong-misconception-dominant profile, with the remainder classified as no-conception or incomplete-data profiles.

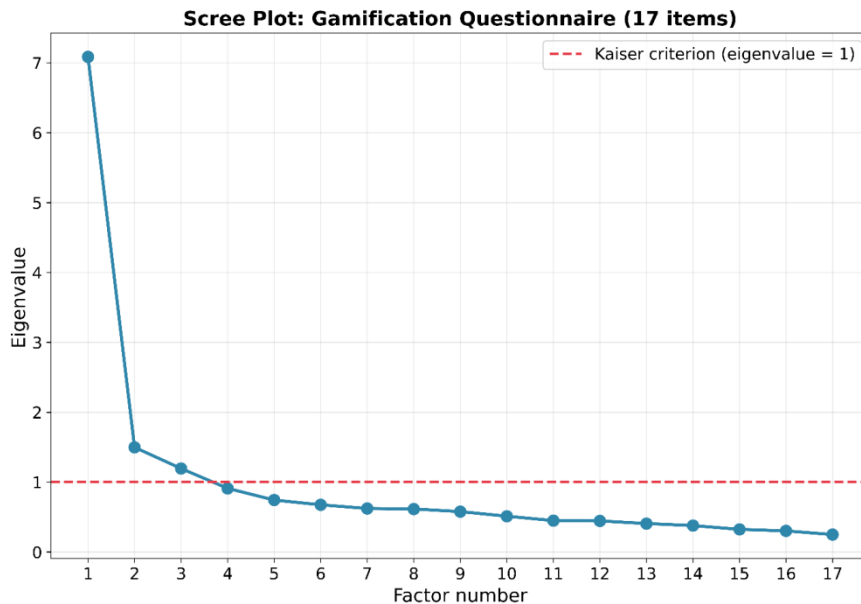
4.2. Factor Structure of the Gamification Instrument (RQ: Structure)

Sampling adequacy was excellent (KMO = .922) and Bartlett's test of sphericity was significant, $\chi^2 = 2547.62$, $p < .001$, confirming that the correlation matrix was suitable for factor analysis (Fabrigar et al., 1999). Three factors exceeded the Kaiser eigenvalue-greater-than-1 criterion, based on the raw (principal-component) correlation matrix (eigenvalues = 7.09, 1.50, 1.19), jointly accounting for 57.5% of total item variance on that basis. When the same three factors were instead extracted as common factors (principal-axis factoring on the reduced, communality-based correlation matrix), they accounted for 48.5% of common variance; the two figures are not directly

comparable because they are computed from different matrices, and both are reported here so that the eigenvalue-based factor count and the variance-explained figure are not read as coming from the same basis. On either basis, the scree plot (see Figure 2) showed a single dominant factor followed by a sharp elbow, which is more consistent with an empirically unidimensional "general gamification engagement" construct than with the four-dimension structure assumed when the instrument was constructed.

Figure 2

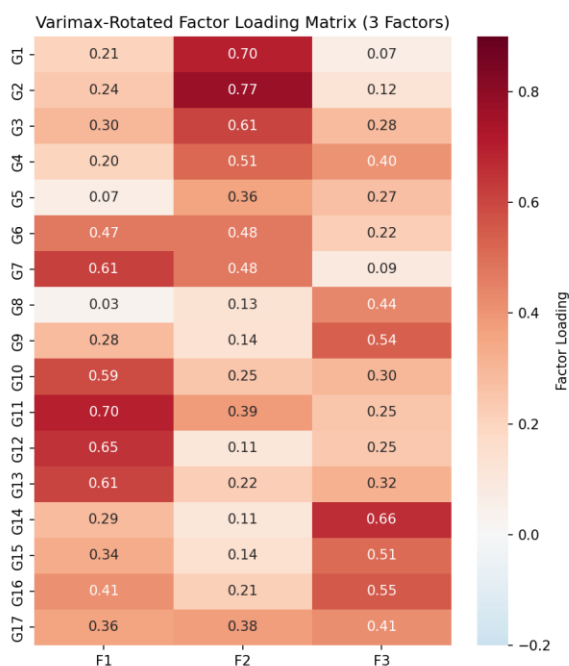
Scree plot of eigenvalues extracted from the 17-item gamification questionnaire



The varimax-rotated three-factor loading matrix (see Figure 3) showed substantial cross-loading on several items (for example, items reflecting engagement and enjoyment loaded appreciably on more than one factor), which further supports the interpretation that the a priori four-dimension model is not strongly recovered in this sample and should be treated as a working, rather than confirmed, structure in subsequent analysis.

Figure 3

Varimax-rotated three-factor loading matrix for the 17 gamification items

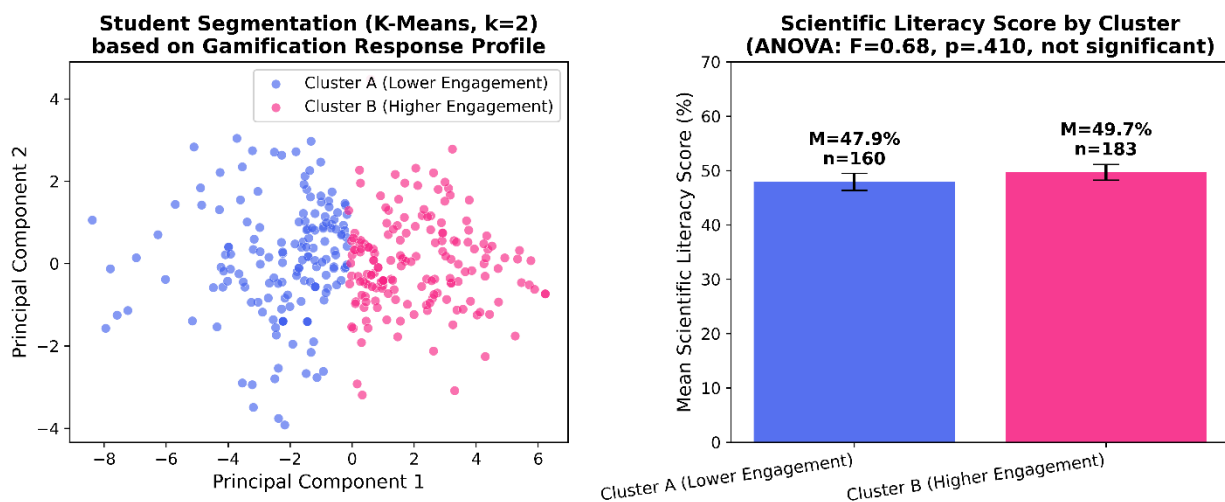


4.3. Student Typology (Cluster Analysis)

K-means clustering achieved its highest average silhouette width at $k = 2$ (silhouette = .223). Following the benchmarks commonly used to interpret this statistic (Rousseeuw, 1987), a value in this range falls below the threshold usually taken to indicate a real, well-separated structure (conventionally around .25-.50 for a "weak, potentially artificial" structure and above .50 for a "reasonable" one); the two-cluster solution is therefore better described as a weakly separated grouping than as clearly distinct profiles, and the H2/H3 tests that follow, which depend on this cluster assignment, should accordingly be treated as exploratory rather than confirmatory. The two clusters corresponded to a lower-engagement profile ($n = 160$, mean total gamification score = 40.3) and a higher-engagement profile ($n = 183$, mean total gamification score = 53.7). As shown in Figure 4, the two clusters were clearly separated on the gamification items themselves but nearly overlapping on the literacy outcome ($M = 47.9\%$ versus 49.7%), a difference that was not statistically significant, $F(1, 341) = 0.68, p = .410$.

Figure 4

Two-cluster K-means segmentation of students by gamification response profile (left) and comparison of mean scientific literacy score between clusters (right).



4.4. Correlation, Network Analysis, and Hypothesis Tests (H1-H3)

Total gamification score and literacy percentage score were essentially uncorrelated under Pearson's r ($r = .084, p = .121$) and only marginally correlated under Spearman's ρ ($\rho = .106, p = .050$), providing no support for H1 as a linear or even a robust monotonic relationship. The network analysis in Figure 5 illustrates this pattern directly: at a threshold of $|r| \geq .30$, all 17 gamification items form a single densely interconnected cluster, while the literacy score node remains completely isolated, sharing no edge above threshold with any gamification item. Item-level correlations with literacy score ranged from $r = .214$ (having previously used gamified media) down to slightly negative values ($r = -.097$ for teacher use of gamification), none reaching a magnitude conventionally considered more than a small effect (Cohen, 1988).

For H2, a chi-square test of independence between the two gamification clusters and the four-category misconception profile was not significant, $\chi^2(3) = 0.55, p = .909$, indicating no association between gamification engagement level and the type of misconception a student was likely to hold. For H3, the association between cluster membership and the five-category ability level was statistically significant, $\chi^2(4) = 20.87, p < .001$, Cramer's $V = .25$, but the effect size falls in the small-to-moderate range and the underlying group means, as already reported, did not differ; the higher-engagement cluster showed a somewhat different distribution across ability categories (proportionally fewer students in the lowest category, more in the middle categories) without a

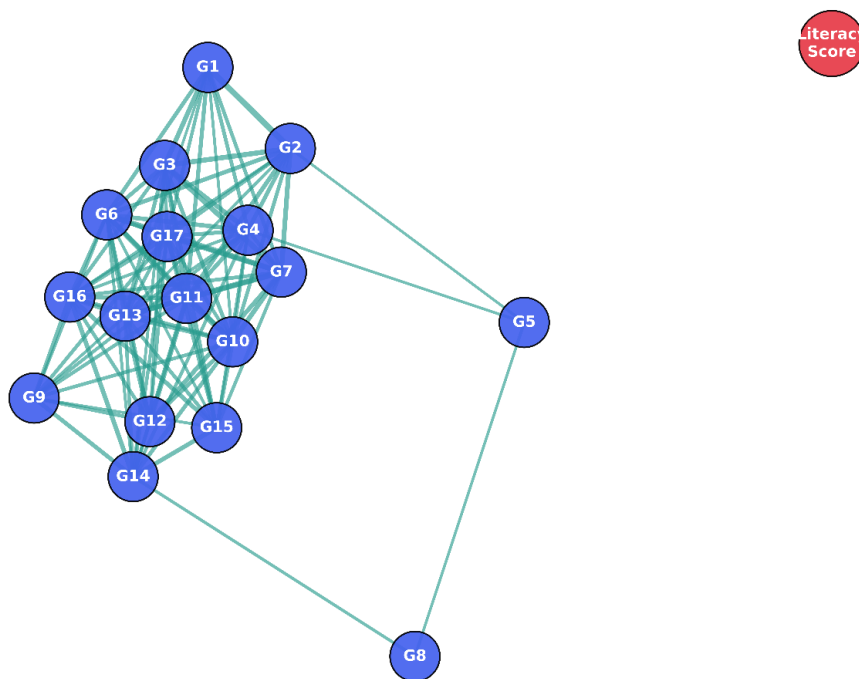
shift in the overall average. H3 therefore receives only partial, distributional rather than mean-level, support.

Figure 5

Correlation network (Spearman $|r| \geq .30$) among the 17 gamification items and the scientific literacy score. The literacy-score node (right) is fully disconnected from the gamification item network

Correlation Network (Spearman $|r| \geq 0.3$) among Gamification Items and Scientific Literacy Score

Red node = Literacy score | green edges = positive correlation | dashed orange = negative



4.5. Structural Equation Model (H1, Path-Level Test)

A structural model specifying four first-order gamification latent factors (Utilization, Engagement, Motivation, Experience) predicting the observed literacy percentage score was estimated. Because the model combines a measurement component (17 indicators loading on four factors) with a structural component (four factors predicting literacy score), fit is reported for the measurement model first and the structural model second, following standard two-step SEM practice (Anderson & Gerbing, 1988). The measurement model alone showed the same fit problem that drives the overall model fit reported below, consistent with the factor-structure result in Section 4.2: forcing four correlated but empirically overlapping factors, one of them (Utilization) reliability-deficient ($\alpha = .44$), into a confirmatory measurement structure produces strain that surfaces as degraded fit. Full-model fit was poor by conventional standards (Hu & Bentler, 1999): comparative fit index (CFI) = .824, Tucker-Lewis index (TLI) = .786, root mean square error of approximation (RMSEA) = .101, $\chi^2/\text{df} = 4.49$. Given this fit, the structural path coefficients below should not be interpreted substantively. None of the four structural paths from gamification factors to literacy score reached significance in the unstandardized solution (Table 5). The very large standard errors relative to the unstandardized coefficients (e.g., Utilization: $b = 45.53$, $SE = 62.51$) indicate estimation instability rather than a precisely estimated near-zero effect, which is expected given the poor measurement model fit and the low reliability of the Utilization factor. The standardized solution, which removes the arbitrary scaling of the latent factors and is the appropriate metric for comparing path magnitudes across the four predictors, is added to Table 5 alongside the unstandardized estimates

and should be read with the same caution given the underlying fit. This pattern is consistent with, rather than an independent confirmation of, the correlational and network findings above, since all three analyses are computed from the same covariance structure among the same variables (see Section 5.1).

Table 5

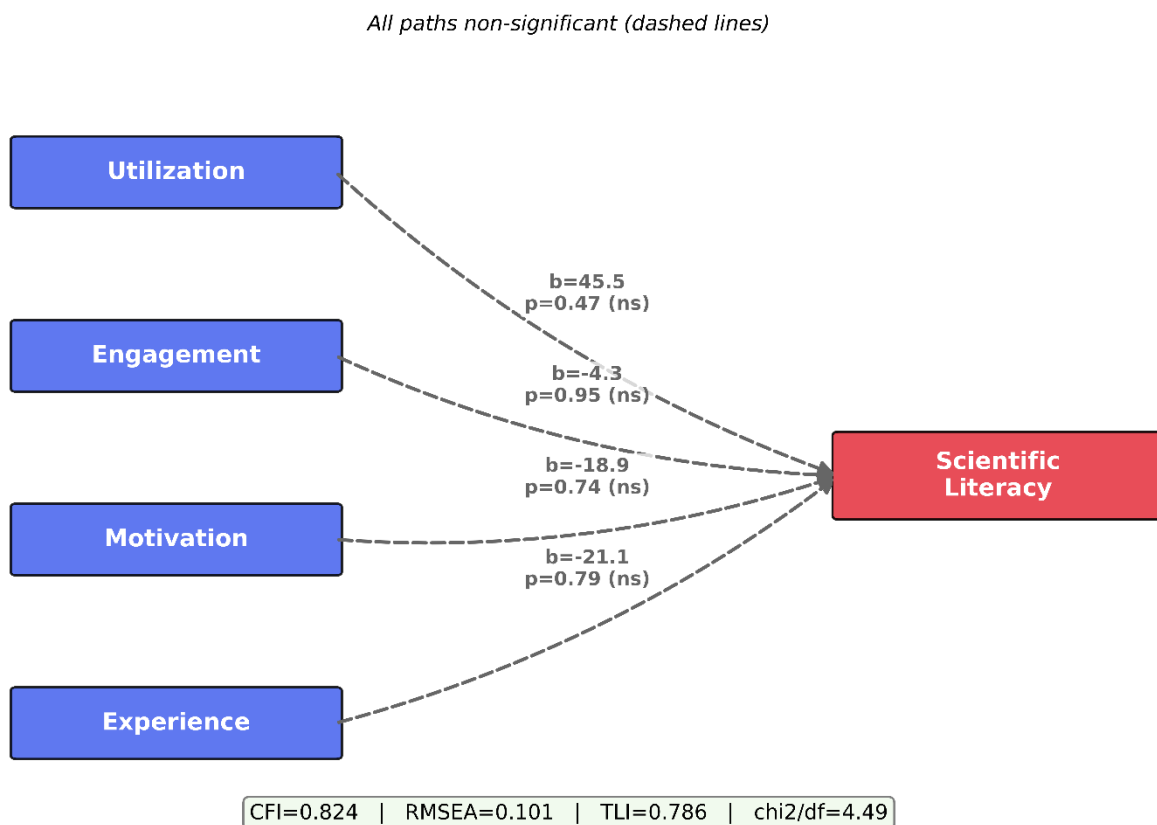
Structural paths from the four gamification factors to the literacy score

Structural path	beta, unstd. (std.)	SE	z	p
Utilization → Literacy score	45.53 (0.08)	62.51	0.73	.466
Engagement → Literacy score	-4.32 (-0.01)	74.29	-0.06	.954
Motivation → Literacy score	-18.92 (-0.04)	57.83	-0.33	.744
Experience → Literacy score	-21.14 (-0.03)	77.91	-0.27	.786

Figure 6

Estimated structural model with four gamification latent factors predicting scientific literacy score.

Structural Model: Four Gamification Dimensions -> Scientific Literacy



Note. All paths are non-significant (dashed lines); model fit indices are reported below the diagram.

4.6. School-Level Moderation (H4)

Because students are nested within eight schools rather than sampled independently, the intraclass correlation coefficient [ICC] for literacy score across schools was estimated from an unconditional (null) multilevel model before testing H4: ICC = .25, indicating that roughly a quarter of the variance in literacy score lies between schools rather than between students within schools. This is a non-trivial degree of clustering, and it means the ordinary least squares regression reported below, which treats the 343 observations as independent, understates the standard errors of any school-related term to some degree; the school main effects and R-squared reported here should therefore be read as approximate rather than exact. An ordinary least squares regression of literacy score on school, total gamification score, and their interaction showed a significant overall model,

$F(15, 327) = 7.90, p < .001, R\text{-squared} = .27$, driven primarily by main effects of school. Critically, none of the seven School $\times$ Gamification-score interaction terms reached significance (all $p > .29$), indicating no detectable evidence that the strength of the gamification-literacy relationship differs across the eight schools in this sample. H4 is therefore not supported, though a multilevel model with a random gamification-score slope by school, appropriately accounting for the nested structure, is recommended as a more rigorous confirmatory test in future work using this or a larger multi-school sample. Taken together, the results across the four analytical strategies applied here converge on the same substantive conclusion: gamification experience, as reported by students in this sample, shows no reliable association with scientific literacy achievement, whether tested at the level of total score, cluster typology, item-level network structure, or latent structural path, and this null pattern does not depend on which school a student attended. As discussed in Section 5.1, these four analyses should be read as convergent descriptions of one underlying null pattern in the same data rather than as fully independent tests.

4.7. PLS-SEM Check of the Structural Result (H1)

The measurement models did not meet the usual PLS-SEM benchmarks (Table 6). In M1, the Utilization construct was unusable: its outer loadings were .945, .181, and $-.213$ for items 1, 5, and 8, its composite reliability was .29, and ρ_A was negative. Engagement (AVE = .44), Experience (.46), and the literacy construct (.31) fell below the .50 benchmark for AVE, and only Motivation (.54) passed it. In M2, the 17-item gamification construct had high reliability ($\alpha = .91, CR = .91$) but an AVE of .38: six of 17 loadings reached .70, and items G5 and G8 (the frequency of game-based lessons and teacher use of games) loaded at .37 and .19. The literacy construct had an AVE of .30 in M2, with one loading at .70 or above (item 9) and three below .40 (items 1, 2, and 3). The test samples different topics, from electron configuration to circular economy, so a single reflective literacy construct fits it poorly. The loadings below are therefore descriptive.

Table 6

Measurement model quality in the PLS-SEM models

Model	Construct	Items	Outer loadings	α	ρ_A	CR	AVE
M1	Utilization	3	$-.213$ to .945	.44	$-.32$	.29	.32
M1	Engagement	5	.330 to .810	.76	.56	.78	.44
M1	Motivation	4	.476 to .816	.75	.76	.82	.54
M1	Experience	5	.392 to .857	.79	.68	.80	.46
M1	Scientific literacy	10	.229 to .734	.74	.77	.81	.31
M2	Gamification	17	.185 to .779	.91	.89	.91	.38
M2	Scientific literacy	10	.192 to .749	.74	.79	.80	.30

Note. M1 = four first-order gamification constructs; M2 = one gamification construct. CR = composite reliability; AVE = average variance extracted. Benchmarks: loadings, α , ρ_A , and $CR \geq .70$; $AVE \geq .50$. The negative ρ_A of Utilization is not a valid reliability estimate.

Discriminant validity failed among the three engagement-related constructs of M1 (Table 7). HTMT was 1.01 for Engagement and Motivation, 1.03 for Engagement and Experience, and 1.03 for Motivation and Experience, and Utilization reached 0.88 to 0.93 against the other three. The square root of AVE was below the inter-construct correlation for all three engagement pairs. Every HTMT involving the literacy construct stayed at or below 0.33. The four gamification dimensions are therefore not empirically distinct in this sample, which repeats the conclusion of Section 4.2, and M2 is the more defensible specification.

In M1, none of the four structural paths was significant by bootstrap (Table 8): Utilization $\beta = .203$ ($p = .144$), Engagement .112 (.302), Motivation $-.001$ (.993), and Experience $-.017$ (.889). R^2 was .073, f^2 was at most 0.03, and inner VIF stayed below 2.9. Figure 7 shows the estimated model.

Table 7

Discriminant validity of the four-construct model (M1)

	Utilization	Engagement	Motivation	Experience	Literacy
Utilization	.57	0.93	0.88	0.89	0.33
Engagement	.53	.66	1.01	1.03	0.22
Motivation	.46	.75	.74	1.03	0.21
Experience	.41	.73	.74	.68	0.17
Literacy	.26	.21	.16	.15	.55

Note. Diagonal: square root of AVE. Below the diagonal: latent variable correlations (Fornell-Larcker). Above the diagonal: HTMT (benchmark < .85 to .90).

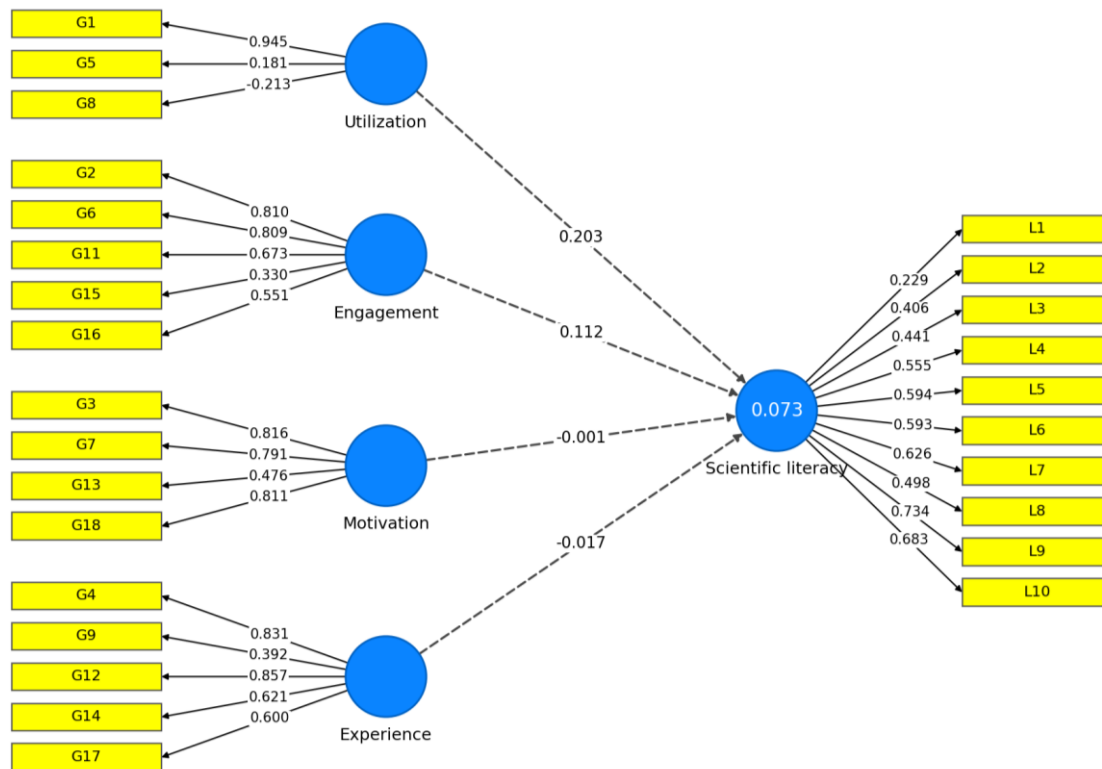
Table 8

Structural results of the PLS-SEM models (bootstrap, 5,000 resamples)

Model	Path	β	SE	<i>t</i>	<i>p</i>	95% CI	f^2	VIF
M1	Utilization → Literacy	.203	.139	1.46	.144	[-.255, .333]	.03	1.41
M1	Engagement → Literacy	.112	.109	1.03	.302	[-.120, .315]	.00	2.89
M1	Motivation → Literacy	-.001	.109	-0.01	.993	[-.187, .230]	.00	2.84
M1	Experience → Literacy	-.017	.120	-0.14	.889	[-.229, .225]	.00	2.63
M2	Gamification → Literacy	.225	.095	2.36	.019	[.190, .358]	.05	1.00

Note. R² of literacy: M1 = .073 (95% CI .062 to .170); M2 = .050 (95% CI .044 to .132). Intervals are percentile intervals.

Figure 7

PLS-SEM estimates for model M1

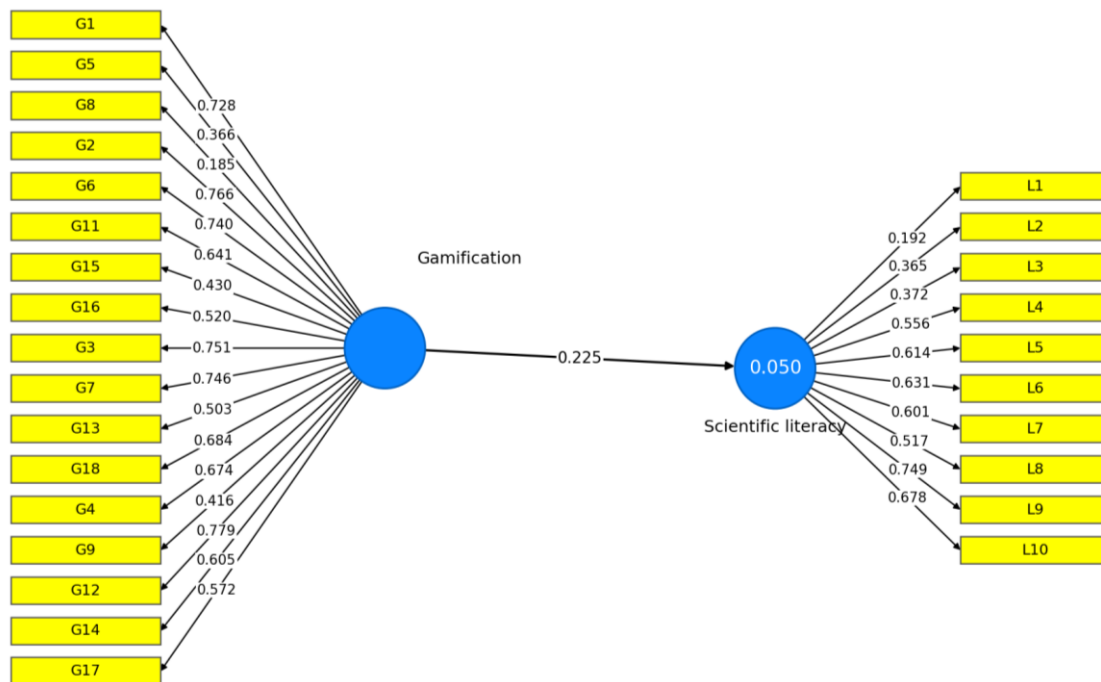
Note. Arrows to indicators carry outer loadings; arrows between constructs carry path coefficients; the value inside the literacy circle is R². Dashed arrows: bootstrap $p \geq .05$.

In M2, the path from gamification to literacy was $\beta = .225$ ($SE = .095$, $p = .019$, 95% CI [.190, .358]), with $R^2 = .050$ and $f^2 = .05$ (Figure 8). This bootstrap p -value overstates the evidence. In the permutation test, R^2 averaged .030 with no association at all, and 5.7% of shuffles produced a coefficient as large as the observed one (permutation $p = .057$). The bootstrap interval also cannot approach zero, because mode A weights make the composite correlate positively with its outcome

by construction. An equal-weight composite of the 17 items correlated $r = .084$ with the literacy percentage ($p = .121$; Section 4.4).

Figure 8

PLS-SEM estimates for model M2 (one gamification construct)



Note. The solid arrow reflects bootstrap $p < .05$; the permutation p was .057.

Predictive relevance was negligible. In M2, Q^2_{predict} ranged from $-.006$ to $.030$ across the ten literacy indicators (M1: $-.013$ to $.038$), so held-out students were predicted no better than by the training mean. The PLS predictions did beat the 17-predictor linear benchmark on nine of ten indicators in M2 and all ten in M1, but that benchmark overfits.

The robustness analyses (see Table 9) produced the same weak pooled estimate with a p -value that depended on the specification. Using the total literacy score as the only indicator gave $\beta = .217$ (bootstrap $p = .065$), and coding incomplete responses as 0 gave $\beta = .224$ ($p = .036$). Centering every indicator within school reduced the estimate to $\beta = .179$ (bootstrap $p = .167$; permutation $p = .514$), and R^2 (.032) fell to the level that shuffled data produce (.034). Adding grade left the gamification path at .176 ($p = .124$). H1 therefore receives weak, unstable support in the pooled sample and none within schools.

Table 9

Robustness of the gamification-literacy association and predictive relevance

Specification	β	Boot. p	Perm. p	R^2	Null R^2	Q^2_{predict} range
M1: four constructs	*	-	.039	.073	.044	-.013 to .038
M2: one construct	.225	.019	.057	.050	.030	-.006 to .030
S1: total score as indicator	.217	.065	.038	.047	.022	.025 to .025
S2: incomplete responses = 0	.224	.036	.051	.050	.030	-.005 to .032
S3: centered within school	.179	.167	.514	.032	.034	-.008 to .017
S4: S3 plus grade	.176	.124	.405	.046	.040	-.011 to .022

Note. β = gamification $\rightarrow$ literacy path. Boot. p = bootstrap p . Null R^2 = mean R^2 under permutation. Perm. p is the share of 5,000 shuffles of the literacy block that produced a coefficient (M2, S1-S4) or R^2 (M1) at least as large as the observed one. S4 also contains a grade indicator; its own path was $\beta = .131$ (bootstrap $p = .143$); *: See Table 8.

5. Discussion

5.1. Interpreting the Predominantly Null Pattern

The findings generally indicate that students' reported gamification experiences were not robustly associated with their scientific literacy performance. Although one of the supplementary analyses suggested a small positive association at the pooled level, this pattern was not sustained when differences between schools were taken into account. Taken together, the results therefore provide limited support for the proposition that greater exposure to gamified learning, as experienced and reported by students, was systematically related to higher scientific literacy in the present context.

This pattern should not be interpreted as evidence that gamification is inherently ineffective. Rather, it may reflect the conditions under which gamification was implemented in the participating schools. Kalogiannakis et al. (2021), in their review of gamification in science education, emphasized that the educational effects of gamification depend substantially on how game elements are designed and integrated into instruction. From this perspective, simply encountering gamified features may not be sufficient to produce measurable gains in learning outcomes. The descriptive findings of the present study suggest that teachers' use of gamified methods was not consistent across classrooms, making it plausible that students experienced different levels, forms, and intensities of gamification. Such variation may weaken the relationship between students' perceived gamification experiences and achievement because the instructional "dose" represented by those experiences is unlikely to be equivalent across classrooms.

This interpretation is also consistent with findings from more structured intervention studies. Núñez-Pacheco et al. (2023) found that engineering students who participated in a deliberately designed gamified intervention achieved higher post-test performance in scientific writing than those receiving conventional instruction, while Okariz et al. (2023) found positive effects of gamification on motivation and the acquisition of skills related to the scientific method in a structured physics laboratory setting. Broader review evidence points in the same direction. In a systematic review of 39 studies, Sökmen et al. (2026) found that technology-enhanced game-based learning environments were generally associated with positive cognitive, behavioral, and affective outcomes, including motivation, engagement, academic achievement, and problem-solving skills. Compared with such interventions, the present study examined naturally occurring differences in students' experiences across existing classroom environments rather than the effects of a standardized gamification treatment. The apparent discrepancy between these findings may therefore be less about whether gamification works and more about the conditions under which its effects become observable. When game elements are systematically embedded within instruction, aligned with learning objectives, and implemented with sufficient consistency, their effects may be more readily detected. In contrast, uneven or relatively limited implementation may generate positive student experiences without producing corresponding differences in measurable academic performance.

The consistency of the overall pattern across the different analytical approaches further supports a cautious interpretation. These analyses should not be regarded as independent replications because several were derived from the same underlying data structure. Their contribution is instead methodological: they indicate that the weak relationship was not simply an artifact of representing gamification through a single summary score or relying on one particular analytical specification. Thus, the present findings narrow the conclusion that can reasonably be drawn. They do not demonstrate that gamification has no educational value; rather, they suggest that student-reported exposure to gamified instruction, when implementation varies across classrooms and schools, may be insufficient on its own to account for meaningful differences in scientific literacy.

5.2. Measurement Structure of the Gamification Instrument

The measurement findings raise an important question about whether gamification experience should be represented as a multidimensional construct in this population. Although the

instrument was originally organized around four dimensions, the empirical structure observed in the present study provided stronger support for a general gamification construct than for clearly differentiated subdimensions. In particular, the instability of the Utilization dimension suggests that interpreting each proposed dimension as a distinct and equally reliable component of gamification may overstate what the instrument can support empirically.

This finding is consistent with established recommendations in exploratory factor analysis that dimensional structures should be determined by the empirical relationships among items rather than assumed from their initial conceptual grouping (Costello & Osborne, 2005; Fabrigar et al., 1999). From this perspective, the original organization of an instrument provides an important theoretical starting point, but it does not by itself establish factorial validity. When the expected dimensions are not clearly reproduced, retaining and interpreting separate subscale scores may introduce construct distinctions that are not sufficiently supported by the observed response patterns. Recent educational scale-development studies similarly treat the proposed theoretical structure as something to be empirically tested rather than assumed. For example, confirmatory analysis of an adapted mathematics-teacher beliefs scale resulted in the removal of six items before an acceptable factor structure was obtained (Çelik Demirci et al., 2023). For the present instrument, treating the items provisionally as indicators of a broader gamification experience construct therefore appears more defensible until the weaker items and dimensions have been revised and the proposed multidimensional structure has been evaluated in an independent sample.

A similar caution has been reported in other educational measurement contexts. Dorsah et al. (2025), in a psychometric re-examination of an epistemic-beliefs instrument among pre-service science teachers, found that theoretically specified dimensions were not fully supported by subsequent confirmatory evidence and that the reliability of some subscales was inadequate. Their findings illustrate a broader measurement issue: dimensions defined during instrument development may not remain equally distinct or stable across samples and populations. The present results extend this concern to the measurement of gamification experience and suggest that theoretically meaningful item groupings should not automatically be treated as empirically established subconstructs.

The findings also clarify the limits of the validation procedures conducted before the main study. Expert review and satisfactory overall internal consistency provide evidence that the items are relevant to the intended domain and that the scale functions coherently at a general level, but neither form of evidence is sufficient to establish multidimensionality. Content validation primarily addresses whether items adequately represent the intended construct, whereas factorial validity concerns whether respondents' answers reproduce the hypothesized internal structure. Similarly, acceptable reliability for the total scale does not ensure that each proposed subscale is itself reliable or psychometrically distinct. The discrepancy between the preliminary validation evidence and the factor-analytic results therefore underscores the importance of examining dimensionality explicitly when an instrument is intended to support interpretations at both total-score and subscale levels.

5.3. Scientific Literacy and Misconception in a Green Chemistry Context

Independent of the gamification question, the descriptive misconception profile obtained through the three-tier diagnostic instrument, with under half of students free of misconception and roughly one in eight showing a strong-misconception-dominant profile on the durian-waste green-chemistry content, is itself a substantive finding worth separate attention. This profile is consistent with the broader three-tier diagnostic literature showing that a meaningful minority of students hold confidently applied incorrect models even after instruction (Caleon & Subramaniam, 2010; Kirbulut & Geban, 2014; Taber, 2002), and it suggests that the durian-waste context, while locally relevant (Do et al., 2025), did not by itself prevent the formation of strong misconceptions. This finding is available for classroom-level remediation regardless of the gamification question and represents an independent contribution of the three-tier instrument developed for this study.

6. Limitations

Several limitations qualify these conclusions. First, the cross-sectional, correlational design cannot rule out that a true but small effect exists and was simply undetected given the observed effect sizes; a properly powered quasi-experimental design with a non-gamified comparison group remains the appropriate next step for a causal claim. Second, gamification exposure was measured through student self-report rather than through independent classroom observation, so the null finding describes the relationship between perceived gamification and achievement, not necessarily between actual implementation fidelity and achievement. Third, the sample was drawn from eight schools within a single city, which limits generalization to other regions or education systems. Fourth, the Utilization subscale's low reliability means that any hypothesis involving that specific dimension should be treated with particular caution until the subscale is revised. Fifth, the scientific literacy measure itself covers a narrow construct scope: it is a ten-item test embedded in a single durian-waste, green-chemistry reading passage, and while it was designed to be a valid, in-depth diagnostic instrument for that specific content, it is not a broad-domain scientific literacy measure in the sense used by large-scale assessments such as PISA, which sample across multiple science topics and competencies. The findings reported here should therefore be understood as applying to scientific literacy as operationalized within this single green-chemistry context, and generalization to scientific literacy more broadly, or to other topic areas, should not be assumed without further validation using a multi-topic instrument. Sixth, the validation evidence for both instruments is limited. Content validity rests on three experts, and with three experts on a four-point scale Aiken's V can take only the values 1.00, .89, .78, and so on, so one rating below 4 moves an item out of the very-valid category. The pilot involved 32 students from two of the eight schools, which is too few for dimension-level reliability or structural analysis of the questionnaire. The literacy test has no construct validity evidence beyond content validation and internal consistency, so its item statistics should be confirmed in a larger sample. Seventh, the schools were visited in sequence over about ten weeks (Table 1), so differences between schools, including the school-level variance reported in Section 4.6, cannot be separated from differences in when each school was tested and where it stood in the school calendar, or from grade level, which was uniform within five of the eight schools (Section 3.4). Eighth, the PLS-SEM check used a Python routine written by the authors rather than SmartPLS. It treated the literacy indicators as 0-3 conception scores because item scores that separate confidence levels are not recoverable, and it rested on measurement models that missed the usual AVE and discriminant validity benchmarks. Its loadings and paths should be read as descriptive and replicated in SmartPLS, ideally with a multilevel specification.

7. Conclusion

This study tested, rather than assumed, the relationship between classroom gamification and scientific literacy achievement among 343 senior high school students in Palu, Central Sulawesi, using a locally grounded, green-chemistry, three-tier diagnostic instrument. Across exploratory factor analysis, cluster analysis, correlation and network analysis, and structural equation modeling, the data converged on a single conclusion: student-reported gamification experience showed no reliable association with scientific literacy achievement or with misconception profile, and only a small, distributional association with ability category that did not correspond to a difference in mean performance. The gamification instrument itself was more consistent with a single general engagement factor than with the four-dimension structure assumed at the outset, and one subscale requires revision before further use. A PLS-SEM check gave a small pooled association ($R^2 = .05$) that was unstable across tests, disappeared within schools, and had no predictive relevance.

8. Implications

8.1. Implications for Classroom Practice

The lowest-rated item in the entire gamification questionnaire concerned the consistency of teacher use of gamified methods, not student attitude toward gamification itself. This points toward implementation fidelity, rather than the mere presence of game elements, as the more promising target for instructional improvement. Teachers and curriculum coordinators in similar settings should treat gamification as a delivery mechanism that requires sustained, systematic use before any effect on achievement can reasonably be expected, rather than as a one-off engagement booster. Independently of the gamification question, the misconception profile obtained from the three-tier diagnostic instrument (Section 4.1) identifies a specific, addressable target for remediation in the durian-waste green-chemistry unit, regardless of whether gamified delivery is used.

8.2. Implications for Measurement and Instrument Design

Researchers adapting or constructing similar gamification questionnaires in comparable Indonesian secondary-school settings should not assume that an a priori multidimensional structure will be recovered empirically; the present instrument requires revision of the Utilization subscale specifically, and re-validation of the four-dimension structure more generally, before it is reused for confirmatory purposes. More broadly, the study illustrates the value of testing instrument structure explicitly with exploratory factor analysis rather than proceeding directly to composite scoring based on the researcher's original item groupings. Expert review and a small pilot are necessary but not sufficient: dimension-level reliability and factor structure need a pilot large enough to estimate them before an instrument enters a main study.

8.3. Implications for Research Design

The consistent null finding obtained here through four complementary, though not statistically independent, analytical strategies argues for the value of multi-method corroboration in educational research generally, and specifically for testing, rather than assuming, both the factor structure of locally adapted instruments and the achievement-level consequences of popular motivational interventions before recommending their broader adoption. Future work in this setting should prioritize a quasi-experimental design with an explicit implementation-fidelity measure, a revised and re-validated gamification instrument, and, where feasible, a multi-site design that allows the moderating role of school-level implementation to be tested with adequate statistical power.

Acknowledgements: The authors would like to express their sincere gratitude to the Ministry of Higher Education, Science, and Technology (Kementerian Pendidikan Tinggi, Sains, dan Teknologi Republik Indonesia) for the research funding support provided through the Directorate of Research, Technology, and Community Service (DPPM) under Contract Number 272/C3/DT.05.00/PL-BARU/2026 dated April 13, 2026.

AI usage and assistance: Statistical analysis (exploratory factor analysis, cluster analysis, correlation and network analysis, structural equation modeling), figure generation, and drafting assistance for this manuscript were supported by Claude (Anthropic). All analytical decisions, interpretation, and final content were reviewed and approved by the authors.

Author contributions: Conceptualization, Dewi Satria Ahmar and Muhammad Fath Azzajjad; Methodology, Dewi Satria Ahmar and Tri Santoso; Investigation, Dewi Satria Ahmar, Muhammad Fath Azzajjad, and Muhammad Ismail; Formal analysis, Dewi Satria Ahmar and Tri Santoso; Writing – original draft, Tri Santoso and Muhammad Fath Azzajjad; Writing – review & editing, Dewi Satria Ahmar and Muhammad Fath Azzajjad.

Data availability: The datasets generated and/or analyzed during the current study are available

from the corresponding author upon reasonable request.

Declaration of interest: The authors declare no conflict of interest.

Ethics declaration: The study was reviewed and approved by the Ethics Committee of Tadulako University, affiliated with LPPM, under approval reference 976/DST/UN28.16/KP.10.00/2026, dated 27 July 2026. The participants were senior high school students, including participants under 18 years of age. Participant confidentiality was protected by removing names prior to statistical analysis, and any identifying information temporarily retained during data collection was subsequently deleted.

Funding: This research was funded by the Ministry of Higher Education, Science, and Technology (Kementerian Pendidikan Tinggi, Sains, dan Teknologi Republik Indonesia) for the research funding support provided through the Directorate of Research, Technology, and Community Service (DPPM) under Contract Number 272/C3/DT.05.00/PL-BARU/2026 dated April 13, 2026.

References

- 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>
- Aikenhead, G. S. (2006). *Science education for everyday life: Evidence-based practice*. Teachers College Press.
- Anderson, J. C., & Gerbing, D. W. (1988). Structural equation modeling in practice: A review and recommended two-step approach. *Psychological Bulletin*, 103(3), 411-423. <https://doi.org/10.1037/0033-2909.103.3.411>
- Arifin, Z., Sukarmin, S., & Saputro, S. (2025). Trends and research frontiers in socioscientific issues for sustainable science education: A systematic and bibliometric analysis from 2014-2024. *Journal of Pedagogical Research*, 9(1), 407-434. <https://doi.org/10.33902/JPR.202530575>
- Bilgin, I., Uzuntiryaki, E., & Geban, O. (2003). Students' misconceptions on the concept of chemical equilibrium. *Journal of Education and Science*, 28(127), 10-15.
- Borsboom, D. (2017). A network theory of mental disorders. *World Psychiatry*, 16(1), 5-13. <https://doi.org/10.1002/wps.20375>
- Byrne, B. M. (2010). *Structural equation modeling with AMOS: Basic concepts, applications, and programming* (2nd ed.). Routledge.
- Caleon, I., & Subramaniam, R. (2010). Development and application of a three-tier diagnostic test to assess secondary students' understanding of waves. *International Journal of Science Education*, 32(7), 939-961. <https://doi.org/10.1080/09500690902890130>
- Chandrasegaran, A. L., Treagust, D. F., & Mocerino, M. (2007). The development of a two-tier multiple-choice diagnostic instrument for evaluating secondary school students' ability to describe and explain chemical reactions using multiple levels of representation. *Chemistry Education Research and Practice*, 8(3), 293-307. <https://doi.org/10.1039/B7RP90006F>
- Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Lawrence Erlbaum Associates.
- Costello, A. B., & Osborne, J. W. (2005). Best practices in exploratory factor analysis: Four recommendations for getting the most from your analysis. *Practical Assessment, Research & Evaluation*, 10(7), 1-9.
- Çelik Demirci, S., Kul, Ü., & Sevimli, E. (2023). Turkish adaptation of the Mathematics Teachers' Beliefs Scale. *Journal of Pedagogical Sociology and Psychology*, 5(2), 92-104. <https://doi.org/10.33902/jpsp.202323416>
- Deci, E. L., Koestner, R., & Ryan, R. M. (1999). A meta-analytic review of experiments examining the effects of extrinsic rewards on intrinsic motivation. *Psychological Bulletin*, 125(6), 627-668. <https://doi.org/10.1037/0033-2909.125.6.627>
- Deci, E. L., & Ryan, R. M. (1985). *Intrinsic motivation and self-determination in human behavior*. Plenum.
- Deterding, S., Dixon, D., Khaled, R., & Nacke, L. (2011). From game design elements to gamefulness: Defining "gamification". In A. Lugmayr, H. Franssila, C. Safran, & I. Hammouda (Eds.), *Proceedings of the 15th International Academic MindTrek Conference: Envisioning Future Media Environments* (pp. 9-15). ACM. <https://doi.org/10.1145/2181037.2181040>
- Dichev, C., & Dicheva, D. (2017). Gamifying education: What is known, what is believed and what remains uncertain: A critical review. *International Journal of Educational Technology in Higher Education*, 14, 9. <https://doi.org/10.1186/s41239-017-0042-5>

- Dijkstra, T. K., & Henseler, J. (2015). Consistent partial least squares path modeling. *MIS Quarterly*, 39(2), 297-316. <https://doi.org/10.25300/MISQ/2015/39.2.02>
- Do, T. H., Nguyen, N. P. N., Duong, T. T. A., Nguyen, D. V., & Vuong, T. X. (2025). Durian peel-seed biochar for efficient methylene blue removal from water: Synthesis, characterization, and adsorption performance. *RSC Advances*, 15(40), 33726-33749. <https://doi.org/10.1039/d5ra05313g>
- Dorsah, P., Amedeker, M. K., & Ngman-Wara, E. I. (2025). Exploring the factor structure of the Epistemic Beliefs Inventory. *International Journal of Didactical Studies*, 6(2), 33004. <https://doi.org/10.33902/ijods.202533004>
- Epskamp, S., Borsboom, D., & Fried, E. I. (2018). Estimating psychological networks and their accuracy: A tutorial paper. *Behavior Research Methods*, 50(1), 195-212. <https://doi.org/10.3758/s13428-017-0862-1>
- Everitt, B. S., Landau, S., Leese, M., & Stahl, D. (2011). *Cluster analysis* (5th ed.). Wiley.
- Fabrigar, L. R., Wegener, D. T., MacCallum, R. C., & Strahan, E. J. (1999). Evaluating the use of exploratory factor analysis in psychological research. *Psychological Methods*, 4(3), 272-299. <https://doi.org/10.1037/1082-989X.4.3.272>
- Fadillah, M. A., Usmeldi, U., Lufri, L., Mawardi, M., & Festiyed, F. (2025). Systematic literature review of inquiry-based learning models for optimizing learning outcomes at various educational levels. *Journal of Pedagogical Sociology and Psychology*, 7(4), 278-293. <https://doi.org/10.33902/jpsp.202532540>
- Field, A. (2013). *Discovering statistics using IBM SPSS statistics* (4th ed.). Sage.
- Fornell, C., & Larcker, D. F. (1981). Evaluating structural equation models with unobservable variables and measurement error. *Journal of Marketing Research*, 18(1), 39-50. <https://doi.org/10.1177/002224378101800104>
- Hair, J. F., Black, W. C., Babin, B. J., & Anderson, R. E. (2010). *Multivariate data analysis* (7th ed.). Prentice Hall.
- Hair, J. F., Hult, G. T. M., Ringle, C. M., & Sarstedt, M. (2022). *A primer on partial least squares structural equation modeling (PLS-SEM)* (3rd ed.). Sage.
- Hamari, J., Koivisto, J., & Sarsa, H. (2014). Does gamification work? A literature review of empirical studies on gamification. In R. H. Sprague Jr. (Ed.), *Proceedings of the 47th Hawaii International Conference on System Sciences* (pp. 3025-3034). IEEE. <https://doi.org/10.1109/HICSS.2014.377>
- Hasan, S., Bagayoko, D., & Kelley, E. L. (1999). Misconceptions and the certainty of response index (CRI). *Physics Education*, 34(5), 294-299. <https://doi.org/10.1088/0031-9120/34/5/304>
- Hellberg, A.-S. (2023). The story of the hatter and the agile methods course: Gamification and game thinking in education. *Journal of Pedagogical Research*, 7(3), 19-42. <https://doi.org/10.33902/JPR.202320130>
- Henseler, J., Ringle, C. M., & Sarstedt, M. (2015). A new criterion for assessing discriminant validity in variance-based structural equation modeling. *Journal of the Academy of Marketing Science*, 43(1), 115-135. <https://doi.org/10.1007/s11747-014-0403-8>
- Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling*, 6(1), 1-55. <https://doi.org/10.1080/10705519909540118>
- Kaiser, H. F. (1974). An index of factorial simplicity. *Psychometrika*, 39(1), 31-36. <https://doi.org/10.1007/BF02291575>
- Kalogiannakis, M., Papadakis, S., & Zourmpakis, A.-I. (2021). Gamification in science education: A systematic review of the literature. *Education Sciences*, 11(1), 22. <https://doi.org/10.3390/educsci11010022>
- Kapp, K. M. (2012). *The gamification of learning and instruction: Game-based methods and strategies for training and education*. Pfeiffer.
- Kirbulut, Z. D., & Geban, O. (2014). Using three-tier diagnostic test to assess students' misconceptions of states of matter. *Eurasia Journal of Mathematics, Science and Technology Education*, 10(5), 509-521. <https://doi.org/10.12973/eurasia.2014.1128a>
- Kline, R. B. (2016). *Principles and practice of structural equation modeling* (4th ed.). Guilford Press.
- Nunez-Pacheco, R., Vidal, E., Castro-Gutierrez, E., Turpo-Gebera, O., Barreda-Parra, A., & Aguaded, I. (2023). Use of a gamified platform to improve scientific writing in engineering students. *Education Sciences*, 13(12), 1164. <https://doi.org/10.3390/educsci13121164>
- Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric theory* (3rd ed.). McGraw-Hill.
- OECD. (2006). *Assessing scientific, reading and mathematical literacy: A framework for PISA 2006*. OECD Publishing. <https://doi.org/10.1787/9789264026407-en>
- OECD. (2023). *PISA 2025 science framework* (second draft). OECD Publishing.

- Okariz, A., Huebra, M., Sarasola, A., Ibarretxe, J., Bidegain, G., & Zubimendi, J. L. (2023). Gamifying physics laboratory work increases motivation and enhances acquisition of the skills required for application of the scientific method. *Education Sciences*, 13(3), 302. <https://doi.org/10.3390/educsci13030302>
- Ratinho, E., & Martins, C. (2023). The role of gamified learning strategies in student's motivation in high school and higher education: A systematic review. *Heliyon*, 9(8), e19033. <https://doi.org/10.1016/j.heliyon.2023.e19033>
- Rousseeuw, P. J. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20(1), 53-65. [https://doi.org/10.1016/0377-0427\(87\)90125-7](https://doi.org/10.1016/0377-0427(87)90125-7)
- Ryan, R. M., & Deci, E. L. (2000). Self-determination theory and the facilitation of intrinsic motivation, social development, and well-being. *American Psychologist*, 55(1), 68-78. <https://doi.org/10.1037/0003-066X.55.1.68>
- Sailer, M., Hense, J. U., Mayr, S. K., & Mandl, H. (2017). How gamification motivates: An experimental study of the effects of specific game design elements on psychological need satisfaction. *Computers in Human Behavior*, 69, 371-380. <https://doi.org/10.1016/j.chb.2016.12.033>
- Seaborn, K., & Fels, D. I. (2015). Gamification in theory and action: A survey. *International Journal of Human-Computer Studies*, 74, 14-31. <https://doi.org/10.1016/j.ijhcs.2014.09.006>
- Sheppard, K. (2006). High school students' understanding of titrations and related acid-base phenomena. *Chemistry Education Research and Practice*, 7(1), 32-45. <https://doi.org/10.1039/B5RP90014J>
- Shmueli, G., Sarstedt, M., Hair, J. F., Cheah, J.-H., Ting, H., Vaithilingam, S., & Ringle, C. M. (2019). Predictive model assessment in PLS-SEM: Guidelines for using PLSpredict. *European Journal of Marketing*, 53(11), 2322-2347. <https://doi.org/10.1108/EJM-02-2019-0189>
- Sökmen, Y., Aygün, E. B., & Ateş Köysu, N. (2026). Integrating augmented reality, virtual reality, and artificial intelligence into game-based learning: A systematic review. *Journal of Pedagogical Sociology and Psychology*, 8(2), e44630. <https://doi.org/10.33902/JPSP.202644630>
- Tabachnick, B. G., Fidell, L. S., & Ullman, J. B. (2019). *Using multivariate statistics* (7th ed.). Pearson.
- Taber, K. S. (2002). *Chemical misconceptions: Prevention, diagnosis and cure* (Vol. 1). Royal Society of Chemistry.
- Treagust, D. F. (1988). Development and use of diagnostic tests to evaluate students' misconceptions in science. *International Journal of Science Education*, 10(2), 159-169. <https://doi.org/10.1080/0950069880100204>
- Vrabec, M., & Proksa, M. (2016). Identifying misconceptions related to chemical bonding concepts in the Slovak school system using the bonding representations inventory as a diagnostic tool. *Journal of Chemical Education*, 93(8), 1364-1370. <https://doi.org/10.1021/acs.jchemed.5b00953>