Journal of Education and Teaching
Vol. 1, No. 1 (2026): P. 15 - 28
https://journal.terbaik.or.id/index.php/JETINDO

Implementing Mystery Box Media in Grade 4 Science: Classroom Action Evidence on Participation and Achievement

Widya Dai[1]

[1]Primary School Teacher Education Study Program, Faculty of Teacher Training and Education, Universitas Pohuwato, Gorontalo, Indonesia

Corresponding author: Widyadai@gmail.com

ABSTRACT

This classroom action study examined the implementation of a low-cost Mystery Box medium and the accompanying changes in participation and achievement in Grade 4 Natural and Social Sciences (IPAS) at a public primary school in Indonesia. The participants were 29 pupils, comprising 10 boys and 19 girls. Conducted from November to December 2025, the study followed a plan-act-observe-reflect sequence across a pre-action stage, two Cycle I meetings, and Cycle II. The intervention combined teacher explanation, concealed instructional prompts, pupil selection and oral response, guided feedback, individual worksheets, and written assessment. Individual score records were reanalyzed using means, standard deviations, medians, ranges, and the proportion reaching the school mastery threshold of 75. Observation records were summarized as 1-4 affective and psychomotor indices, while a 10-item teacher-activity checklist described implementation. Mean scores were 53.79 (SD = 22.74) at pretest, 70.93 (SD = 15.99) at Cycle I Meeting 1, 68.97 (SD = 28.58) at Cycle I Meeting 2, and 88.97 (SD = 13.72) at Cycle II. Mastery increased monotonically from 24.14% to 41.38%, 58.62%, and 89.66%, exceeding the 80% classroom success criterion. Teacher-activity completion rose from 5 of 10 to 8 of 10 and then 10 of 10 items, while both observation indices increased. From baseline to Cycle II, 24 pupils improved their recorded scores and 19 newly crossed the mastery threshold. The evidence supports local classroom improvement when the Mystery Box is used as a structured prompt-response-feedback routine, not merely as a novel object. Because there was no comparison group, the assessments were not documented as psychometrically equivalent, and no delayed test was administered, the findings do not establish causal efficacy or durable learning.

Keywords: classroom action research; integrated science and social studies; learning achievement; Mystery Box; primary education; student participation

1. INTRODUCTION

Indonesia's primary curriculum combines natural and social science within Ilmu Pengetahuan Alam dan Sosial (IPAS). For Phase B, which includes Grades 3 and 4, the curriculum expects pupils to identify changes in the state of matter in daily life and to develop process skills such as observing, questioning, predicting, analyzing simple data, reflecting, and communicating findings (Ministry of Primary and Secondary Education,

E-ISSN: xxxx-xxxx
P-ISSN: xxxx-xxxx
DOI :

n.d.). These expectations require more than the delivery of factual content. They require classroom structures that make pupils attend, respond, explain, and receive feedback.

The classroom problem addressed in this study was practical and local. Initial observation and a teacher interview at SDN 02 Tilamuta indicated that Grade 4 IPAS lessons relied mainly on teacher explanation and a textbook. Pupils were described as having difficulty answering questions, explaining the material, sustaining attention, and participating actively. The formal pretest later confirmed a low level of mastery: only 7 of 29 pupils reached the school's minimum criterion of 75. The instructional challenge was therefore not simply to add a visually attractive resource, but to reorganize how pupils encountered and responded to science content.

A Mystery Box is a three-dimensional classroom medium that conceals content, objects, images, or questions until a learner selects or reveals them. Its instructional promise rests on a plausible information-gap mechanism. When learners recognize that relevant information is missing but obtainable, the gap can direct attention and information seeking (Loewenstein, 1994). In science education, curiosity is most defensibly understood through observable exploration, information seeking, and question-related behavior rather than through a general claim that an activity is fun (Jirout & Klahr, 2012; Jirout, 2020). The box may therefore create an entry point for curiosity, but curiosity must not be inferred merely from the presence of a concealed object.

The medium may also alter the mode of participation. Selecting a prompt and producing an answer requires more overt engagement than listening to an explanation. The ICAP framework distinguishes passive, active, constructive, and interactive engagement, with stronger learning expected when pupils generate or exchange ideas rather than only receive information (Chi & Wylie, 2014). At the same time, young learners need guidance. Meta-analytic evidence on inquiry-based learning indicates that support improves learning, especially when task demands exceed learners' independent capacity (Lazonder & Harmsen, 2016). A Mystery Box lesson is therefore better conceptualized as guided active participation than as unguided discovery.

Recent Indonesian studies provide encouraging but still bounded evidence. Development studies have reported that Mystery Box products can be feasible and accompany improved science and IPAS outcomes (Alina & Fathurrahman, 2024; Nuradhisti & Prasetyaningtyas, 2025). Quasi-experimental studies have also reported higher motivation, achievement, or interest among elementary pupils using related media (Nuhaa & Witanto, 2024; Syaira & Meilana, 2025). These findings justify further investigation, but they do not show that every box design works in the same way. The quality of the prompts, teacher guidance, opportunities to respond, and assessment practices may be more consequential than the container itself.

A second gap concerns implementation. Endpoint comparisons can obscure whether instruction was delivered consistently, whether participation changed before achievement, and whether the score distribution developed uniformly. Action research is appropriate when the aim is to improve a bounded practice through cycles of planning, action, observation, and reflection (Kemmis et al., 2014). Its value lies in local problem solving and traceable learning about practice. It should not be presented as an efficacy trial. Reviews of science education action research and teacher research emphasize the need for transparent cycles, reflexivity, evidence quality, and explicit limits on inference (Laudonia et al., 2018; Leuverink & Aarts, 2019).

This study therefore asked: (1) How did recorded implementation and pupil participation develop across the classroom action sequence? (2) How did achievement scores and mastery rates change from the pre-action stage through Cycle II? (3) What can, and cannot, be concluded from the convergence of implementation, observation, and achievement records? The article contributes a reanalysis of student-level data, corrects inconsistencies in the thesis summary tables, and distinguishes local improvement from causal effectiveness.

2. LITERATURE REVIEW

2.1 Mystery as a Pedagogical Affordance

The term Mystery Box describes a material format, not a complete pedagogy. Concealment can focus attention because learners anticipate a reveal, but novelty alone is temporary. Loewenstein's (1994) information-gap account suggests that curiosity is activated when a learner becomes aware of a specific gap in knowledge and expects that the gap can be closed. In a classroom, this means that mystery must be connected to an intelligible question. If the concealed content is unrelated, too difficult, or revealed without requiring thought, the box may entertain without supporting conceptual learning.

For children, scientific curiosity involves a willingness to approach uncertainty and seek explanatory information (Jirout & Klahr, 2012). Jirout (2020) further argues that educators can support early scientific thinking by creating manageable uncertainty and opportunities for question asking and exploration. The intervention in this study did not measure curiosity directly. Consequently, curiosity is treated as a design affordance that may have helped pupils enter the task, while observed participation and written scores remain the actual evidence.

2.2 Guided Participation, Response, and Feedback

Engagement is multidimensional. Behavioral participation, emotional involvement, and cognitive investment should not be collapsed into a single label (Fredricks et al., 2004). The source records categorized pupils separately in affective and psychomotor domains, but they did not provide validated measures of emotional or cognitive engagement. These observations can describe classroom movement, attention, and willingness to participate, but they cannot establish a latent engagement construct.

E-ISSN: xxxx-xxxx
P-ISSN: xxxx-xxxx
DOI :

The Mystery Box sequence nevertheless changes what pupils are asked to do. They select a concealed question, retrieve relevant content, state an answer, and receive teacher guidance before completing an individual task. In ICAP terms, these actions can move a lesson beyond passive reception toward active or constructive engagement (Chi & Wylie, 2014). The extent of that movement depends on the quality of the response. Choosing a slip is only active manipulation; explaining why ice melts or why water freezes is more constructive because the pupil produces a relation or explanation.

Guidance is equally important. Evidence across inquiry-learning studies indicates that prompts, scaffolds, and teacher support generally strengthen learning (Lazonder & Harmsen, 2016). Formative assessment theory also positions elicitation and feedback as part of an instructional process in which evidence of learning is used to decide the next step (Black & Wiliam, 2009). A Mystery Box is most likely to be educationally useful when the reveal is followed by an interpretable response, feedback, and an opportunity to correct or elaborate understanding.

2.3 Interpreting Improvement in Classroom Action Research

Classroom action research aims to improve practice within a specific site. Its cyclical logic allows teachers and researchers to identify a concern, implement an action, observe what changes, and revise practice (Kemmis et al., 2014). In science education, this design can connect research directly to instruction, but the quality of the account depends on how clearly the action, adaptations, evidence, and reflection are documented (Laudonia et al., 2018).

Without a comparison condition, repeated assessment, maturation, increased familiarity with the material, teacher attention, and changes in implementation can all contribute to an upward trajectory. The defensible claim is therefore that improvement occurred during the action sequence and was consistent with the intended instructional logic. A claim that the medium caused the improvement would require stronger control over alternative explanations and better evidence that the assessments were comparable across occasions.

3. METHOD

3.1 Design, Setting, and Participants

The study used classroom action research organized around the recurring phases of planning, action, observation, and reflection. The source thesis described two cycles, with two recorded meetings in Cycle I and one final Cycle II assessment. The study was conducted from November to December 2025 at SDN 02 Tilamuta, a public primary school in Boalemo Regency, Gorontalo Province, Indonesia. The pretest was administered on November 13, 2025, and the first Cycle I action was recorded on November 18, 2025. Later meeting dates were not specified in the archived narrative.

All 29 pupils in one Grade 4 class participated, including 10 boys and 19 girls. The research was conducted collaboratively with the classroom teacher, although the archive did not consistently distinguish which instructional components were led by the teacher and which were led by the student researcher. The unit addressed changes in the state of matter, including freezing, melting, evaporation, condensation, and related everyday examples.

3.2 The Mystery Box Action

The intervention used a decorated three-dimensional box containing instructional content and rolled question prompts. The recorded lesson sequence included an opening activity, teacher explanation of the target concept, introduction of the Mystery Box, pupil selection of a concealed prompt, oral response, opportunities for questions, reinforcement or feedback, an individual worksheet, and a written end-of-meeting assessment. The box therefore served as an organizing device for attention and response rather than as a stand-alone source of information.

Reflection records stated that the next meeting was intended to address weaknesses identified in the previous meeting. However, the archive repeated the same broad lesson steps and did not specify the exact instructional modifications between meetings. Exact dosage was also inconsistent across the planning documents and is not treated as an analytic variable. Table 1 presents only the action and evidence that can be reconstructed with confidence.

Table 1

Reconstructed Classroom Action Sequence and Evidence

Stage	Recorded classroom action	Primary evidence
Pre-action	Teacher-led explanation and textbook use; no Mystery Box; baseline written test	Teacher interview, pupil observation, individual pretest scores
Cycle I, Meeting 1	Mystery Box presentation, concealed questions, oral response, worksheet, written test	10-item teacher-activity checklist, pupil observation, individual scores
Cycle I, Meeting 2	Continuation after reflection using the same documented sequence; exact adaptation not specified	10-item teacher-activity checklist, pupil observation, individual scores
Cycle II	Continuation after Cycle I reflection; guided Mystery Box response and final assessment	10-item teacher-activity checklist, pupil observation, individual scores

Note. The source archive did not specify the precise modifications made after each reflection or provide a consistent record of lesson duration.

3.3 Data Sources and Measures

Four data sources were available: a baseline teacher interview, structured classroom observation, written achievement records, and lesson documentation. The

E-ISSN: xxxx-xxxx
P-ISSN: xxxx-xxxx
DOI :

Journal of Education and Teaching
Vol.1, No.1 Mei (2026): start page-end page

interview was used only to characterize the initial instructional problem. No verbatim transcript was available, so the article reports the teacher's documented responses as summaries and does not present invented quotations.

Achievement was represented by individual scores on classroom assessments converted to a 0-100 scale. The school mastery criterion was 75, and the action-research success criterion was that at least 80% of the class reach mastery. The methods section of the thesis described five-item objective tests. However, Cycle I Meeting 1 scores were recorded in increments of approximately 16-17 points (50, 66, 83, and 100), whereas the other occasions mainly used 20-point increments. The assessments were therefore not assumed to be psychometrically equated.

Pupil observations placed each pupil in one of four categories for an affective domain and one of four categories for a psychomotor domain. The scale table defined 4 as very good, 3 as good, 2 as adequate, and 1 as poor. Category counts were converted to a weighted class mean ranging from 1 to 4. A contradictory legend below the Cycle II table reversed the scale, but the table headings and all earlier scale documentation retained the 4-to-1 order; the repeated scale definition was used for the reanalysis. These indices are descriptive observations, not validated measures of affect or psychomotor competence.

Instructional implementation was documented with a 10-item yes/no teacher-activity checklist covering opening activities, learning objectives, use of the Mystery Box, content explanation, classroom management, communication, reinforcement, worksheet guidance, reflection, and closure. The number marked yes was used as a simple implementation count. The archive did not report interrater reliability or independent observation.

3.4 Data Analysis and Integrity Checks

All achievement summaries were recalculated from the 29 pupil-level records. For each occasion, the analysis computed the mean, sample standard deviation, median, range, number reaching 75, and mastery percentage. Baseline-to-Cycle II change was also examined at the pupil level by counting higher, unchanged, or lower scores and by cross-classifying baseline and final mastery status. Percentage-point change was used for the class mastery rate. No hypothesis test or causal effect size was calculated because the study involved one intact class, no comparison condition, non-equated classroom assessments, and a locally defined stopping criterion.

The reanalysis identified an error in the thesis summary table: the Cycle I Meeting 1 and Meeting 2 means and mastery percentages had been transposed. The individual records show a Meeting 1 mean of 70.93 with 12 pupils reaching mastery (41.38%), and a Meeting 2 mean of 68.97 with 17 pupils reaching mastery (58.62%). These corrected values are used throughout this article.

One Cycle I Meeting 2 score was recorded as zero. The archive did not indicate whether this represented absence, noncompletion, or performance. The value was retained as recorded, and no imputation or sensitivity estimate was presented as the primary result. This decision preserves the source data but contributes to the wide score dispersion at that meeting.

4. FINDINGS

4.1 Baseline Classroom Condition

The teacher's documented baseline responses described instruction dominated by explanation and textbook use, limited variation in media, difficulty attracting pupil attention, and low pupil participation. These responses do not independently establish the cause of low achievement, but they identify the practical concern that guided the action: pupils had few structured opportunities to reveal and discuss their understanding during the lesson.

The pretest mean was 53.79 (SD = 22.74), with a median of 60 and scores ranging from 20 to 100. Seven pupils reached the mastery criterion and 22 did not, producing a class mastery rate of 24.14%. The affective observation index was 1.83 and the psychomotor observation index was 1.93 on the 1-4 descriptive scale. The baseline therefore combined low class mastery with observation ratings centered near the adequate-to-poor range.

4.2 Development of Instructional Implementation and Participation

The teacher-activity checklist showed a clear implementation trajectory. At Cycle I Meeting 1, 5 of 10 items were marked yes. The uncompleted items concerned explicit communication of objectives, content mastery, classroom management, reinforcement, and pupil reflection. At Cycle I Meeting 2, 8 of 10 items were completed; classroom management and reflection remained unmarked. By Cycle II, all 10 items were marked yes. Thus, the intervention did not enter the classroom as a fully stable treatment. Its recorded implementation became more complete over time.

The pupil observation indices increased in the same sequence. The affective index rose from 1.83 at baseline to 2.17, 2.45, and 2.66. The psychomotor index rose from 1.93 to 2.21, 2.62, and 2.86. The pattern is gradual rather than abrupt. It suggests that more pupils were classified in the good or very good categories as the action progressed, although the absence of indicator definitions and reliability evidence means the indices should be read as structured classroom impressions.

Table 2

Implementation and Pupil Observation Trajectory

Stage	Teacher-activity items completed	Affective index (1-4)	Psychomotor index (1-4)
Pre-action	Not applicable	1.83	1.93
Cycle I, Meeting 1	5 of 10 (50%)	2.17	2.21

E-ISSN: xxxx-xxxx
P-ISSN: xxxx-xxxx
DOI :

Cycle I, Meeting 2	8 of 10 (80%)	2.45	2.62
Cycle II	10 of 10 (100%)	2.66	2.86

Note. Observation indices are weighted means of category counts using 4 = very good, 3 = good, 2 = adequate, and 1 = poor. They are not standardized scales.

Convergence between the teacher checklist and pupil observations is analytically important. The final stage combined complete recorded implementation with the highest observation indices. The data cannot determine which checklist component mattered most, but they challenge a simple novelty explanation. If the box alone had been sufficient, a large and stable change would be expected at its first use. Instead, participation indicators strengthened as the broader instructional sequence became more complete.

4.3 Achievement and Mastery Across the Action Sequence

Achievement improved overall, but the mean trajectory was not strictly linear. After the pretest mean of 53.79, the mean rose to 70.93 at Cycle I Meeting 1, then declined slightly to 68.97 at Meeting 2 before reaching 88.97 at Cycle II. The Meeting 2 standard deviation increased to 28.58 because the distribution included both high scores and one recorded zero. By Cycle II, the standard deviation had fallen to 13.72 and the median was 100.

Table 3

Recalculated Achievement and Mastery by Stage

Stage	Mean	SD	Median	Range	Pupils at or above 75, n (%)
Pretest	53.79	22.74	60	20-100	7 (24.14%)
Cycle I, Meeting 1	70.93	15.99	66	50-100	12 (41.38%)
Cycle I, Meeting 2	68.97	28.58	80	0-100	17 (58.62%)
Cycle II	88.97	13.72	100	60-100	26 (89.66%)

Note. Values were recalculated from the individual score records. The zero at Cycle I Meeting 2 was retained as recorded. The assessments were not documented as psychometrically equivalent across stages.

Unlike the mean, mastery increased at every recorded stage: 24.14%, 41.38%, 58.62%, and 89.66%. The final rate exceeded the action-research success criterion of 80% by 9.66 percentage points. Relative to baseline, class mastery increased by 65.52 percentage points, and the mean score increased by 35.17 points. These values describe the recorded class, not a population parameter.

The divergence between the Cycle I mean and mastery rate is informative. Meeting 2 had a lower mean than Meeting 1 but a higher median and more pupils above 75. Threshold attainment and average performance therefore captured different features of the distribution. Reporting only the class mean would hide the larger number of pupils crossing the mastery threshold, while reporting only mastery would hide the dispersion and the recorded zero.

4.4 Pupil-Level Change from Baseline to Cycle II

At the pupil level, 24 of 29 pupils recorded a higher score at Cycle II than at baseline, five recorded the same score, and none recorded a lower final score. The median individual increase was 40 points. These descriptive changes were not uniform: some pupils gained 20 points, while others gained as much as 80 points. Five pupils showed no raw-score change, including pupils who were already at mastery and pupils who remained below it.

Table 4

Baseline-to-Cycle II Mastery Transitions

Transition	Number of pupils	Percentage of class
Maintained mastery	7	24.14%
Newly reached mastery	19	65.52%
Remained below mastery	3	10.34%
Moved below mastery	0	0.00%

Note. Mastery was defined as a recorded score of at least 75. Percentages use N = 29.

All seven pupils who met mastery at baseline retained it at Cycle II, and 19 of the 22 initially below mastery crossed the threshold. Three pupils remained below, each with a final score of 60. The stopping criterion was therefore met, but the remaining cases show that class-level success did not eliminate the need for differentiated follow-up.

4.5 Integrated Reading of the Findings

Three strands developed together: implementation checklist completion, observation indices, and class mastery. Their convergence is consistent with the proposition that a more fully implemented prompt-response-feedback sequence supported classroom participation and achievement. The evidence is stronger than a simple before-after percentage because it shows process and distribution, but it remains observational within one action sequence.

The strongest finding is therefore not that a decorated box independently produced an effect. It is that the class reached its locally defined success criterion while the recorded instructional routine became more complete, participation indices improved, and most pupils crossed the mastery threshold. Alternative explanations remain possible, including repeated exposure, teacher adaptation, practice with the content, and increased familiarity with classroom assessment.

E-ISSN: xxxx-xxxx
P-ISSN: xxxx-xxxx
DOI :

5. DISCUSSION

5.1 The Instructional Function Was More Important Than the Object

The findings support an interpretation of the Mystery Box as an instructional organizer. Concealment may have created a manageable information gap, but the learning-relevant work occurred after the reveal: pupils had to retrieve knowledge, state a response, listen to feedback, and complete an individual task. This distinction matters because it prevents novelty from being mistaken for pedagogy. Loewenstein's (1994) account explains why a specific unknown can focus attention, while Jirout and Klahr (2012) show why curiosity should be connected to information seeking rather than inferred from excitement alone.

The observed classroom sequence also illustrates the boundary between active and constructive engagement. Physically drawing a prompt is active, but explaining a change of state or relating it to daily experience requires learners to generate meaning. This distinction is consistent with ICAP (Chi & Wylie, 2014). The present records do not show the quality of each oral explanation, so they cannot establish that every pupil reached a constructive mode. They do show that the lesson provided more occasions for overt response than the baseline textbook-and-explanation format.

The intervention was guided rather than inquiry-based in a full sense. Pupils responded to teacher-prepared questions; they were not documented as generating investigable questions, designing investigations, or collecting evidence. This qualification is important for international readers. The pedagogical strength lies in scaffolded participation and formative elicitation, not in a claim that the box constituted open scientific inquiry.

5.2 Implementation Quality and Participation Developed Together

The teacher-activity checklist suggests that implementation quality was initially partial. The first meeting lacked several components that can influence learning, including clear objectives, classroom management, reinforcement, and reflection. Completion rose to 8 of 10 and then 10 of 10 items. The Mystery Box was therefore embedded in a changing instructional environment rather than delivered as a fixed treatment.

This finding aligns with broader evidence in primary science. Bennett et al. (2023) found that promising approaches often depend on opportunities for pupils to make thinking visible through talk, receive feedback, and work within supported classroom interaction. Black and Wiliam's (2009) formative assessment framework similarly emphasizes the use of elicited evidence to guide subsequent instruction. In the present class, concealed questions gave the teacher a public view of what pupils could retrieve, while the worksheet and written assessment provided individual evidence.

The rising observation indices are consistent with this implementation account, but they should not be overinterpreted. Affective and psychomotor labels are broad, and the source did not define their indicators or establish reliability. Following Fredricks et al. (2004), observable participation should not be equated automatically with deep cognitive engagement. Future studies should specify behaviors, use more than one observer, and distinguish attention, emotional response, oral contribution, and conceptual explanation.

5.3 Achievement Improved, but the Trajectory Was Not Uniform

The final achievement pattern is educationally meaningful within the class. The mean rose by 35.17 points, 26 pupils reached mastery, and 19 crossed the threshold after being below it at baseline. The class exceeded its predefined 80% criterion. These figures justify concluding that the action sequence met its local improvement goal.

However, the Cycle I results demonstrate why stronger analysis requires more than a single upward line. The mean declined slightly from Meeting 1 to Meeting 2 even as the mastery rate increased. The Meeting 2 distribution had a higher median, more pupils at or above 75, and substantially greater dispersion. One zero score amplified that dispersion. If it represented absence rather than performance, the mean would not describe the same participation base. Because the archive is silent, retaining the value and reporting the ambiguity is more defensible than silently changing it.

Assessment comparability is a second qualification. The score increments indicate that Cycle I Meeting 1 may have used a different number of items from the other occasions, despite the methods section referring generally to five-item tests. Differences in item number, difficulty, and content coverage can produce score changes independent of learning. The result is best described as a trajectory across classroom assessments, not as a standardized gain estimate.

5.4 Relationship to Previous Mystery Box Studies

The final pattern is broadly consistent with Indonesian studies reporting improved elementary outcomes or interest with Mystery Box media (Alina & Fathurrahman, 2024; Nuhaa & Witanto, 2024; Nuradhisti & Prasetyaningtyas, 2025; Syaira & Meilana, 2025). The present study adds a process-oriented qualification. It shows that implementation became more complete across meetings and that the relationship between mean score and mastery was not uniform. The contribution is therefore not another unqualified effectiveness claim, but a classroom-level explanation of how the medium was embedded in an evolving routine.

The study also narrows the pedagogical claim. A Mystery Box is likely to be useful when it contains curriculum-aligned prompts, makes every pupil responsible for a response, supports explanation, and feeds evidence back into instruction. A product that is attractive but cognitively thin may not produce the same pattern. This interpretation is compatible with the guided-learning evidence summarized by Lazonder and Harmsen (2016).

E-ISSN: xxxx-xxxx
P-ISSN: xxxx-xxxx
DOI :

5.5 Practical Implications

For teachers, the practical unit of design should be the full response cycle. The box should pose one clear, curriculum-aligned uncertainty; pupils should predict or answer before the reveal; the teacher should ask for a brief reason; feedback should address the concept rather than only correctness; and every pupil should complete an individual check. This sequence keeps the mystery subordinate to the learning goal.

For monitoring, teachers should record both average performance and threshold attainment. The present Cycle I data show that these indicators can move differently. A short record of noncompletion, absence, and item difficulty would prevent ambiguous values from distorting interpretation. Teachers should also document precisely what was changed after reflection, such as prompt wording, wait time, grouping, feedback, or support for pupils who remained below mastery.

For the three pupils who remained below mastery, repeating the same whole-class activity is not automatically sufficient. Follow-up should diagnose whether the difficulty concerns vocabulary, the sequence of state changes, reading the prompt, attention, or task completion. The Mystery Box can then be adapted with fewer choices, visual representations, worked examples, paired explanation, or additional guided practice.

5.6 Limitations and Directions for Research

The study has several limitations. First, one class, one school, and no comparison condition prevent causal attribution and statistical generalization. Second, the assessment forms were short and not documented as equivalent in content, difficulty, or reliability. Third, no delayed assessment established whether the final pattern persisted after the novelty and repeated practice ended.

Fourth, implementation roles, lesson dosage, and cycle-specific adaptations were not documented consistently. Fifth, the teacher-activity checklist and pupil observation categories did not include interrater reliability or validated operational definitions. Sixth, the study did not directly measure curiosity, cognitive engagement, or the quality of pupils' explanations. Claims about mechanism are therefore interpretive and should be tested rather than treated as established.

Future action studies should use parallel or equated assessments, retain item-level responses, distinguish absence from zero performance, document every instructional adaptation, and include a delayed test. Video or dual-observer coding could improve implementation and participation evidence. A validated curiosity or situational interest measure, combined with coded question asking and explanation quality, could test whether the information-gap mechanism actually mediates learning.

Stronger comparative research could assign or match classes to a Mystery Box sequence and a content-equivalent non-mystery sequence. This comparison would help separate concealment and anticipation from the effects of additional questions, feedback, practice, and teacher attention. Multi-site studies could then examine whether outcomes vary by teacher preparation, class size, prior attainment, or the complexity of the science topic.

6. CONCLUSION

The Grade 4 class met its local action-research goal during the implementation of Mystery Box media. Mastery increased from 24.14% at baseline to 89.66% at Cycle II, the final mean reached 88.97, and 19 pupils newly crossed the mastery threshold. Teacher-activity checklist completion and pupil observation indices also increased. The most defensible interpretation is that improvement occurred as the box became part of a more complete prompt-response-feedback-assessment routine.

The study does not demonstrate that novelty or the physical box independently caused achievement. Its value lies in showing a feasible, low-cost classroom action and in making the limits of the evidence explicit. Future research should retain the practical strengths of the routine while using comparable assessments, precise implementation records, direct measures of curiosity and engagement, and designs that can address causality and retention.

AUTHOR CONTRIBUTIONS

Proposed CRediT statement for author confirmation. Widya Dai: conceptualization, methodology, investigation, data curation, formal analysis, visualization, writing the original draft, and review and editing. The author must confirm that these roles accurately describe the work before submission.

FUNDING

DATA AVAILABILITY

De-identified aggregate and pupil-level score records supporting the descriptive findings may be available from the corresponding author on reasonable request, subject to school permission, participant confidentiality, and applicable ethical restrictions.

GENERATIVE ARTIFICIAL INTELLIGENCE DISCLOSURE

OpenAI's ChatGPT was used to support English-language editing, manuscript restructuring, recalculation of descriptive statistics from the thesis tables, reference verification, and document formatting. It was not used to create participants, observations, scores, interview quotations, or research procedures. The author remains responsible for verifying the data, references, ethics statement, analysis, and final submitted manuscript.

E-ISSN: xxxx-xxxx
P-ISSN: xxxx-xxxx
DOI :

REFERENCES

Alina, K. F., & Fathurrahman, M. (2024). Pekendo Mystery Box media to improve fourth grade science and social learning outcomes for elementary school students. Jurnal Penelitian dan Pengembangan Pendidikan, 8(3), 528-537. https://doi.org/10.23887/jppp.v8i3.79076

Bennett, J., Dunlop, L., Atkinson, L., Compton, S., Glasspoole-Bird, H., Lubben, F., Reiss, M. J., & Turkenburg-van Diepen, M. (2023). A systematic review of approaches to primary science teaching. Education Endowment Foundation. https://educationendowmentfoundation.org.uk/education-evidence/evidence-reviews/primary-science

Black, P., & Wiliam, D. (2009). Developing the theory of formative assessment. Educational Assessment, Evaluation and Accountability, 21(1), 5-31. https://doi.org/10.1007/s11092-008-9068-5

Chi, M. T. H., & Wylie, R. (2014). The ICAP framework: Linking cognitive engagement to active learning outcomes. Educational Psychologist, 49(4), 219-243. https://doi.org/10.1080/00461520.2014.965823

Fredricks, J. A., Blumenfeld, P. C., & Paris, A. H. (2004). School engagement: Potential of the concept, state of the evidence. Review of Educational Research, 74(1), 59-109. https://doi.org/10.3102/00346543074001059

Jirout, J. J. (2020). Supporting early scientific thinking through curiosity. Frontiers in Psychology, 11, Article 1717. https://doi.org/10.3389/fpsyg.2020.01717

Jirout, J., & Klahr, D. (2012). Children's scientific curiosity: In search of an operational definition of an elusive concept. Developmental Review, 32(2), 125-160. https://doi.org/10.1016/j.dr.2012.04.002

Kemmis, S., McTaggart, R., & Nixon, R. (2014). The action research planner: Doing critical participatory action research. Springer. https://doi.org/10.1007/978-981-4560-67-2

Laudonia, I., Mamlok-Naaman, R., Abels, S., & Eilks, I. (2018). Action research in science education: An analytical review of the literature. Educational Action Research, 26(3), 480-495. https://doi.org/10.1080/09650792.2017.1358198

Lazonder, A. W., & Harmsen, R. (2016). Meta-analysis of inquiry-based learning: Effects of guidance. Review of Educational Research, 86(3), 681-718. https://doi.org/10.3102/0034654315627366

Leuverink, K. R., & Aarts, A. M. L. (2019). A quality assessment of teacher research. Educational Action Research, 27(5), 758-777. https://doi.org/10.1080/09650792.2018.1535445

Loewenstein, G. (1994). The psychology of curiosity: A review and reinterpretation. Psychological Bulletin, 116(1), 75-98. https://doi.org/10.1037/0033-2909.116.1.75

Ministry of Primary and Secondary Education. (n.d.). Capaian pembelajaran: Ilmu Pengetahuan Alam dan Sosial (IPAS), Fase B [Learning outcomes: Natural and Social Sciences, Phase B]. Retrieved July 27, 2026, from https://guru.kemendikdasmen.go.id/kurikulum/referensi-penerapan/capaian-pembelajaran/sd-sma/ilmu-pengetahuan-alam-dan-sosial-ipas/fase-b/

Nuhaa, F. N., & Witanto, Y. (2024). Mystery Box media increases motivation and learning achievement of fourth grade students in civics learning. Jurnal Penelitian dan Pengembangan Pendidikan, 8(3), 538-546. https://doi.org/10.23887/jppp.v8i3.79420

Nuradhisti, S. A., & Prasetyaningtyas, F. D. (2025). Development of Mystery Box learning media to improve elementary school students' learning outcomes. Jurnal Penelitian Pendidikan IPA, 11(3), 907-916. https://doi.org/10.29303/jppipa.v11i3.10801

Syaira, S., & Meilana, S. F. (2025). The effect of Mystery Box media on fourth grade students' interest in science learning. Journal of Educational Sciences, 9(5), 3308-3321. https://doi.org/10.31258/jes.9.5.p.3308-3321