

Beyond knowledge gains: remembering, reconstructing, caring, and the limits of action in animation-based industrial heritage education

Xin Bian

Shandong Women's University

Andre Brown

andre.brown@vuw.ac.nz

Victoria University of Wellington

Bruno Marques

Victoria University of Wellington

Article

Keywords: Anthropocene, animation, environmental learning, industrial heritage, place-based education

Posted Date: October 22nd, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-10945174/v1>

License:   This work is licensed under a Creative Commons Attribution 4.0 International License. [Read Full License](#)

Additional Declarations: No competing interests reported.

Abstract

Industrial heritage makes the environmental consequences and social dependencies of technological development tangible, but making such histories memorable does not establish critical understanding or action. This mixed-methods evaluation examined a five-minute, place-based animation among 452 Year 1–6 students in China. Evidence comprised four knowledge assessments over two months, learning-experience questionnaires, 111 drawings, and interviews with 11 teachers and four heritage guides. Baseline-adjusted knowledge differences favoured the animation-supported condition at two-week and two-month follow-ups, although performance declined after the immediate post-test in both conditions. Questionnaires and interviews indicated positive learning experiences and stewardship intentions, whereas drawings showed uneven reconstruction of narrative and causal relationships. Evidence of action-oriented understanding remained limited. The findings position animation as an accessible entry point to extractive histories while showing why retention, represented understanding, expressed care, and demonstrated action should not be treated as interchangeable educational outcomes.

1. Introduction

Environmental and sustainability education cannot be judged only by what learners remember; it must also consider how they interpret contested human–environment relations and what forms of responsibility or action become imaginable (UNESCO 2020). For younger learners, place-based education can make such relations intelligible by connecting environmental change with familiar landscapes, community knowledge, and lived experience (Li and Shein 2023). Yet local relevance does not itself produce critical understanding. Critical place-based education asks how places are produced, transformed, remembered, and contested, shifting attention from learning in a place to critically reading the social and ecological relations through which that place has been made (Häggström and Schmidt 2020; Tugurian and Carrier 2017).

Industrial heritage makes this educational tension especially visible because material remains connect technological and economic development with labour histories, community dependence, and collective memory (Morgado, Licursi, and Da Silva 2024). Viewed through an Anthropocene lens, such landscapes do more than illustrate environmental change: they reveal how present conditions have been historically produced through uneven relationships among extraction, technology, labour, communities, and land (Garsdal and Nordgren 2026). Mining landscapes also embody environmental damage and restoration, making the distribution of development benefits and costs available for inquiry (Author(s) 2026a). Approached as a social-ecological text rather than only as a collection of old structures or technical facts, industrial heritage can therefore locate contemporary sustainability questions within the histories that produced them (Cuenca-López, Martín-Cáceres, and Estepa-Giménez 2021). For primary learners, the challenge is to coordinate historical sequence, spatial transformation, causal relationships, and competing values within an intelligible account (Li and Shein 2023). Industrial heritage should consequently do more than provide an attractive local backdrop; it should enable examination of relationships among landscapes, labour, communities, and ecological systems (Häggström and Schmidt 2020).

Making this critical potential accessible requires a medium that can represent long-term change without dissolving its causal and ethical complexity. Animation can coordinate visual and verbal information to make temporal sequences and spatial transformation perceptually available (Mayer 2024). Yet animation is transient, and fluent viewing may support recognition without requiring learners to construct an explanation (Schubertová et al. 2023). Visibility must therefore be distinguished from causal understanding. Storytelling performs another function by organising events, people, time, place, and values into a coherent account (Robin 2008). Evidence that participatory, learner-authored

storying can support ecological imagination and agency cannot be transferred directly to a professionally produced, teacher-mediated animation (Kumpulainen et al. 2023).

These representational distinctions create a corresponding evaluation problem. Environmental education evaluations often report favourable knowledge, attitude, or engagement outcomes more readily than they identify which programme features produced them (Stern, Powell, and Hill 2014). They may also allow one positive indicator to stand for a broader educational transformation. In this study, remembering, reconstructing, caring, and acting are treated as related but analytically distinct domains. Knowledge retention cannot establish independent causal reconstruction, while interest or stewardship intention cannot establish action competence (Oinonen et al. 2024). Evaluating animation-supported learning therefore requires evidence sources that retain their own inferential limits before their findings are compared.

Together, these conceptual and methodological problems define a focused research gap. Existing evaluations rarely establish whether the accessibility and engagement afforded by a short narrative animation are accompanied by durable remembering, independent reconstruction of temporal and causal relationships, or action-oriented understanding. The present study addresses this gap by comparing animation-supported and conventional instruction across four knowledge-test occasions and by integrating learning-experience measures, student drawings, and teacher and heritage-guide accounts. Its purpose is to determine what kinds of environmental learning the evaluated instructional arrangement supports and where claims based on each evidence source reach their limits.

To translate this gap into an evaluative inquiry while keeping conclusions aligned with the limits of each evidence source, the study was guided by three questions. (1) Under matched lesson time and content, how did animation-supported and conventional instruction differ in immediate learning experience and in knowledge-test trajectories from T0 to T3? (2) What did student drawings and teacher and heritage-guide accounts reveal about narrative reconstruction, classroom usability, social-ecological meaning, and implementation limits? (3) Where did quantitative and interpretive evidence converge or diverge across remembering, reconstructing, caring, and acting? The longitudinal test assessed maintenance on repeated items; it did not establish behavioural transfer or action competence.

The study contributes to environmental education research in two related ways. Conceptually, it treats remembering, reconstructing, caring, and acting as connected but non-substitutable outcomes, preventing favourable knowledge or attitude indicators from representing educational transformation as a whole. Methodologically, it integrates longitudinal, visual, questionnaire, and interview evidence without treating any source as a criterion measure for another. This evidence architecture shifts evaluation from asking simply whether animation works to examining what it enables, what it leaves unresolved, and what further pedagogy is required.

2. Literature review

2.1. Industrial heritage as critical place-based environmental learning

Place-based education locates environmental meaning in relationships among local environments, social practices, and lived experience (Russ et al. 2015). Local relevance can support belonging and ecological meaning-making, but critical place-based education goes further by asking how places are produced, remembered, governed, and contested (Häggström and Schmidt 2020; Riveiro-Rodríguez et al. 2021; Li and Shein 2023). Place is therefore not

merely a familiar setting for environmental content. It is an object of inquiry through which learners can examine power, memory, access, and responsibility.

Industrial heritage makes these relationships unusually visible because one landscape may embody technological achievement, livelihood, displacement, pollution, identity, and repair. This historical depth is central to education in the Anthropocene: present environmental conditions are not isolated problems but outcomes of earlier relationships among extraction, technology, institutions, communities, and land (Garsdal and Nordgren 2026). It also guards against reducing the Anthropocene to an undifferentiated account of human impact by keeping unequal histories, responsibilities, and capacities for action in view. Heritage education can connect material remains with historical interpretation and citizenship (Cuenca-López, Martín-Cáceres, and Estepa-Giménez 2021; Morgado, Licursi, and Da Silva 2024), while community research shows that heritage meanings are negotiated rather than simply transmitted (Li et al. 2020). Environmental learning must consequently move beyond identifying industrial features to explaining how extraction and technological change reorganised relationships among people, land, resources, and ecological systems.

For primary learners, this move requires the coordination of historical sequence, spatial transformation, causal explanation, and value conflict. Familiarity or affection may strengthen environmental identity, but neither demonstrates understanding of contested development choices (Tugurian and Carrier 2017). Educational resources must therefore reduce avoidable complexity without removing the social-ecological trade-offs that make industrial heritage significant. The central representational problem is how to make long-term transformation accessible while retaining enough causal and ethical complexity for critical interpretation.

2.2. Animation, narrative, and differentiated environmental learning outcomes

Animation can make temporal and spatial change perceptually available by coordinating visual and verbal information. Multimedia learning theory explains how learners select, organise, and integrate information under limited working-memory capacity (Mayer 2024). Yet educational video effects depend on design and instructional use rather than on the medium alone (Tani, Manuguerra, and Khan 2022). Transient information and fluent viewing may support recognition without requiring learners to construct causal explanations, and animation does not invariably outperform less dynamic materials (Schubertová et al. 2023). Visibility must therefore be distinguished from reconstruction.

Narrative contributes continuity by organising people, events, time, place, and values into a meaningful sequence (Lambert 2013; Robin 2008). Participatory storying can support ecological imagination and agency when learners contribute perspectives and negotiate meaning (Häggström 2022; Kumpulainen et al. 2023; Tzima et al. 2020). A professionally produced, teacher-mediated animation offers a different educational experience. Digital heritage research similarly suggests that visual simulation can facilitate access while remaining insufficient for deeper historical and cultural interpretation unless context and meaning are deliberately designed into the experience (Xie et al. 2026). It can orient attention and emotion, but viewing cannot be equated with authoring, deliberating, or acting. Digital media should thus be assessed for what they make accessible and what forms of participation they leave unavailable (Lowan-Trudeau 2023).

This distinction separates four related but non-substitutable domains. Remembering concerns retention of presented content; reconstructing concerns independent organisation of temporal, spatial, and causal relations; caring includes interest, emotional response, social-cultural awareness, and stewardship intention; acting concerns knowledge of possible actions, reasoned choice, confidence, and participation. Action competence therefore exceeds willingness to

protect the environment (Sass et al. 2020; Koskela and Paloniemi 2023; Oinonen et al. 2024). Longitudinal evidence also cautions against inferring action competence from short-term attitudes (Olsson, Gericke, and Boeve-de Pauw 2022).

The distinction also determines what programme evaluation can claim. Tests can document performance on specified content, questionnaires can capture reported experience and intention, drawings can reveal selected temporal, spatial, and causal relations, and interviews can clarify implementation and perceived meaning. No source establishes the others. Environmental education evaluation is therefore strongest when it connects purpose, implementation, and outcomes, preserves the inferential limits of each source, and examines convergence and divergence across remembering, reconstructing, caring, and acting rather than aggregating favourable indicators into a general claim of effectiveness (Stern, Powell, and Hill 2014; Garip et al. 2021; Gu et al. 2026).

3. Materials and methods

3.1. Evaluation overview and evidence architecture

Figure 1 distinguishes prior program-development studies from evidence analysed in this article. Row 1 contains earlier literature, stakeholder, needs, and heritage-value studies used only to establish context and design provenance. Rows 2–3 contain the present comparative outcomes and interpretive evidence. Each source was analysed separately before integration.

Integration examined convergence, divergence, alternative explanations, and evidential strength across remembering, reconstructing, caring, and acting. These are interpretive domains, not chronological stages or interchangeable measures. Table 1 states the function and inferential boundary of each source.

Table 1
Evidence sources, evaluation functions, and inferential boundaries.

Evidence source	Evaluation purpose	Participants / material	Principal contribution	Inferential boundary
Program context and learner needs	Establish the educational problem and intended learners.	Literature synthesis; 33 stakeholder interviews; questionnaire from 399 Year 1–6 students.	Limited prior exposure alongside high curiosity, willingness to learn, and support for an age-accessible resource.	Documents need and design rationale; does not estimate intervention outcomes.
Place and content development	Translate Dagushan industrial heritage into a coherent social–ecological narrative.	Archival, spatial, and interview evidence; heritage-value hierarchy weighted by 131 experts.	Historical and sociocultural value formed the narrative core; technological and ecological evidence supplied explanatory content.	Supports content provenance and credibility, not learning effectiveness.
Quantitative outcome evidence	Estimate comparative immediate outcomes and maintenance of test performance.	452 linked students; questionnaire at T0–T1; knowledge test at T0, T1, T2, and T3.	Adjusted group differences favoured the animation-supported condition and persisted through T3 while attenuating.	The design does not isolate animation from the wider instructional arrangement or demonstrate transfer.
Drawing and interview evidence	Interpret open-ended reconstruction, implementation, credibility, and program limitations.	111 eligible drawings; 11 teachers; 4 heritage guides.	Revealed non-linear grade variation, strong accessibility evidence, and weak action-oriented evidence.	Explains and qualifies outcomes; does not independently validate the quantitative effect.
<i>Note.</i> Evidence sources were analysed according to their own constructs and then integrated. Questionnaire and knowledge-test estimates, drawing codes, and interview accounts were not treated as interchangeable measures.				

3.2. Program setting and development evidence

Row 1 draws entirely on related prior studies: a review of industrial-heritage digital storytelling (Author(s) 2026b), stakeholder consultation, a 399-student needs survey and media-preference analysis (Author(s) 2025b), animation development and early user testing (Author(s) 2025a), and a heritage-value study weighted by 131 experts (Author(s) 2026a). These studies established the program’s need, content, and design constraints; their data were not reanalysed as outcomes or treated as independent confirmation of the present 452-student evaluation.

Prior stakeholder interviews addressed curriculum provision, educational value, communication, media preferences, and recommendations; the student needs survey assessed awareness, curiosity, willingness to learn, perceived values, and media preferences. These background studies used age-appropriate paper or online administration and established design rationale rather than intervention effectiveness.

3.3. Outcome evaluation design, participants, and procedure

The outcome sample was formed by random selection from the complete student cohort within each of Years 1–6 at one Qingdao primary school, rather than by selecting intact classes. Of 470 returned records, 18 substantially incomplete or patterned records were excluded, leaving 452 cases (96.17%): 226 in each instructional condition, distributed across Years 1–6 (72, 68, 88, 68, 73, and 83 students). Random sampling within grades did not constitute

random assignment to instructional condition, and individual random assignment was not documented. Participation was voluntary, consented, anonymised, and unrelated to grading. Ethical approval and consent details are provided in the declarations section.

Both groups completed baseline measures (T0). The animation-supported group viewed the 5 min 02 s animation with brief teacher clarification; the comparison group received matched oral, textbook, and board-based instruction. Both completed the immediate post-test (T1), a two-week knowledge follow-up (T2), and a fixed two-month follow-up (T3). The four datasets contained the same 452 anonymised IDs, invariant condition labels, no missing total scores, and item totals consistent with the scoring key.

3.4. Quantitative measures and analysis

3.4.1. Learning-experience questionnaire

The learning-experience questionnaire contained eight five-point dimensions: Interest, Comprehension, Knowledge Acquisition, Emotional Response, Educational Value, Engagement Level, Social and Cultural Awareness, and Heritage Stewardship Intention (Cronbach's $\alpha = .92$; CFA: $\chi^2/df = 1.033$, RMSEA = .009, CFI = .997). It was administered at T0 and T1. For each dimension, the T1 score was regressed on condition, the corresponding T0 score, and grade indicators using HC3 standard errors. Holm adjustment was applied across the eight outcomes. Individual T0-T1 gains and standardised between-condition effect sizes were also reported. These models estimated comparative program-level differences while accounting for baseline score and grade.

3.4.2. Knowledge test and longitudinal analysis

The same seven-item knowledge test (0–7 correct) assessed historical sequence, mine depth, mineral resources, and heritage listing at T0-T3. A linear mixed-effects model included condition, categorical time, their interaction, and participant random intercept. Baseline-adjusted follow-up differences and within-condition declines were reported with 95% confidence intervals, Holm adjustment, and standardised effects. Repeated items measured maintenance of familiar performance, not transfer, and the quasi-experimental design does not isolate animation as the causal mechanism.

3.5. Interpretive evidence

3.5.1. Student drawings

The submitted archive contained 120 drawing files. SHA-256 screening identified five duplicate groups; four drawing identities appeared under conflicting grade labels and were excluded, while one within-grade duplicate was counted once. The final corpus comprised 111 grade-attributable drawings: Year 1, 23; Year 2, 17; Year 3, 16; Year 4, 18; Year 5, 17; and Year 6, 20. All eligible drawings were retained irrespective of artistic quality or agreement with the expected narrative. A frozen codebook operationalised nine ordinal dimensions covering narrative, temporal, causal, spatial, schematic, value-related, environmental, industrial-heritage, and stewardship features. Drawings were treated as visual interpretive evidence with ordinal coding, not as a second estimate of the questionnaire or knowledge-test effect.

Two trained coders independently scored all nine drawing dimensions before reconciliation. Because the initial industrial-heritage (HER) scores differed systematically, a third trained coder independently rescored HER for the complete archive using clarified anchors. The three-person median represented HER, while the other eight dimensions used the mean of the two full coders. Reliability was assessed with two-way absolute-agreement ICCs, quadratic-weighted Cohen's κ , exact agreement, and agreement within one scale point. Five grade-blinded machine-assisted

passes provided an exploratory grade-level sensitivity analysis reported in Fig. 5 and Table 7. They were not treated as a gold standard or as a substitute for researcher coding. Archive processing and drawing analyses used Python 3.12.13, Pillow 12.3.0, pandas 3.0.5, NumPy 2.5.1, SciPy 1.18.0, and Matplotlib 3.11.1.

3.5.2. Teacher and heritage-guide interviews

The interpretive strand examined implementation, perceived meaning, and open-ended reconstruction after the animation had been produced. Participants comprised 11 primary teachers (A01–A11) across Mathematics, Chinese, English, Science, and Art, 4 industrial-heritage guides (A12–A15), and the children represented by the drawing archive. Seven teachers contributed to both pre-design and post-view rounds, while four contributed only to the first round. Post-view interviews lasted 10–22 minutes and produced transcripts of 1,761–4,795 words. Teachers and guides viewed the animation before being interviewed through Zoom, Tencent Meeting, or WeChat; interviews were recorded and transcribed with consent.

The teacher guide contained 20 questions across nine domains: educational effectiveness, primary-school suitability, visual and auditory experience, emotional resonance, pacing, interest, student competencies, responsibility, and improvement. Teachers were asked to distinguish directly observed student responses from their own professional judgments. The heritage-guide guide addressed historical accuracy, interpretation, sustainability, emotional resonance, and comparison with other media. Transcripts were analysed in NVivo 12 through open, axial, and selective coding, producing 38 open categories, eight axial dimensions, and three higher-order categories. Excerpts were then classified as direct observation, teacher-reported student speech, professional judgment, or recommendation. Theme frequencies represent coded mentions, not participant prevalence. A second researcher reviewed the coding; most nodes had Cohen's $\kappa > .61$ and more than 75% exceeded .75.

3.6. Cross-source integration and inferential boundaries

Questionnaire, test, drawing, and interview data were analysed separately before a joint display compared convergence, divergence, and evidential boundaries across remembering, reconstructing, caring, and acting. No source served as a criterion measure for another, preventing self-report or test gains from being treated as behavioural transfer or action competence.

4. Results

4.1. Program context and design provenance

The prior focused review located 19 industrial-heritage digital-storytelling publications from 2011 to February 2026, mostly museum or prototype studies and only two educational applications (Author(s) 2026b). This background finding motivated, but was not part of, the outcome evaluation.

Stakeholder interviews established expert consensus: 30 of 33 participants (90.9%) considered primary-level industrial heritage education necessary. Of the 27 who expressed a medium preference, 23 (85.2%) recommended animation, the highest of any option, followed by storytelling (70.4%). These findings indicate both strong support for the educational goal and convergent endorsement of animation as the appropriate medium.

The prior needs survey ($n = 399$) found low awareness (35.84%) and understanding of industrial buildings (37.84%) despite high curiosity (68.67%) and willingness to learn (86.72%). Animation was the preferred medium and was positively associated with comprehension ($r = .248$, $p < .01$; Author(s) 2025b). These data informed design but were not included in the present outcome sample.

A related site study combined archival, spatial, interview, and expert evidence into a seven-dimension heritage-value hierarchy; historical (.24) and social-cultural (.20) values ranked highest (Author(s) 2026a). Together with the needs evidence, this informed the three-generation, 107-year animation narrative previously described in detail (Author(s) 2025a).

Prior teacher accounts identified shallow curriculum coverage, rare site visits, uneven technology, and limited teacher knowledge. Recommendations for a short, age-appropriate, locally connected story with interactive follow-up explain the program format; they do not demonstrate effectiveness.

4.2. Comparative outcomes

4.2.1. Learning-experience outcomes

With design credibility and implementation fit established through formative evaluation, the analysis now turns to summative outcome judgement. The summative evaluation included 452 students randomly sampled from the full student cohort within each of Years 1–6, rather than selected by intact class; 226 were in the experimental group and 226 in the control group. The sample covered all six grades with a relatively balanced distribution: Grade 1, 72 students (15.93%); Grade 2, 68 students (15.04%); Grade 3, 88 students (19.47%); Grade 4, 68 students (15.04%); Grade 5, 73 students (16.15%); and Grade 6, 83 students (18.36%). Grade composition was comparable between groups (Table 2).

Before the intervention, both groups completed the experiential questionnaire measuring eight dimensions, namely Interest, Comprehension, Knowledge Acquisition, Emotional Response, Educational Value, Engagement Level, Social and Cultural Awareness, and Heritage Stewardship Intention. Independent-samples t-tests on pre-test scores showed no significant differences between groups on any dimension, with p values ranging from 0.088 to 0.836 and mean differences between 0.007 and 0.062 points on the five-point scale (Table 2). Pre-test knowledge-test accuracy was similar between groups (experimental = 31.9%, control = 29.6%; see Content knowledge). These results indicate that the experimental and control groups started from a broadly similar baseline and that initial differences are unlikely to account substantially for post-test outcomes.

Table 2
Measured baseline characteristics of the two instructional conditions (pre-test).

Variable	Experimental (N = 226) Mean ± SD	Control (N = 226) Mean ± SD	p
Interest	2.84 ± 0.40	2.80 ± 0.45	.308
Comprehension	2.88 ± 0.40	2.82 ± 0.37	.088
Knowledge Acquisition	2.79 ± 0.38	2.78 ± 0.42	.697
Emotional Response	2.84 ± 0.41	2.78 ± 0.36	.096
Educational Value	2.78 ± 0.42	2.81 ± 0.41	.451
Engagement Level	2.77 ± 0.39	2.82 ± 0.40	.190
Social and Cultural Awareness	2.80 ± 0.41	2.79 ± 0.35	.836
Heritage Stewardship Intention	2.81 ± 0.43	2.79 ± 0.44	.611
<i>Note. p values were obtained from independent-samples t-tests. No statistically significant baseline differences were detected on the measured variables; however, unmeasured baseline differences cannot be excluded because assignment to instructional condition was not documented as random.</i>			

After adjustment for the corresponding baseline score and grade, post-test scores favoured the animation-supported condition across all eight learning-experience outcomes. Adjusted differences between the animation-supported and comparison conditions ranged from 0.65 points for educational value to 0.77 points for emotional response on the five-point scales; all Holm-adjusted p values were below 0.001 (Fig. 2).

The observed post-test differences ranged from 0.655 to 0.767 points, with an average difference of approximately 0.72 points on the five-point scale. The largest gaps involved emotional response, social-cultural awareness, and perceived knowledge acquisition. These outcomes identify the dimensions on which the instructional conditions differed most strongly; they do not establish why the differences occurred.

Both groups improved from pre-test to post-test across the eight learning-experience outcomes, while the animation group showed consistently larger mean gains. Figure 3 reports absolute change on the original five-point scales and standardised between-group differences in change. The effect-size estimates are descriptive comparative evidence and should not be interpreted as effects of random assignment.

Note

Violin plots show the distributions of individual pre–post change scores (post-test minus pre-test) for the experimental (blue) and control (grey) groups. Embedded boxplots indicate the median and interquartile range (IQR), and black dots represent the mean change scores. Positive values indicate improvements from pre-test to post-test. Panels represent (a) composite learning experience, (b) interest, (c) comprehension, (d) knowledge acquisition, (e) emotional response, (f) educational value, (g) engagement, (h) social and cultural awareness, and (i) heritage stewardship intention.

All eight experiential dimensions in the animation group increased from pre-test to post-test ($p < .001$; Table 3). Mean change scores ranged from 1.009 to 1.141, corresponding to relative gains of approximately 35.1–40.8%, with an average improvement of 38.9%. The largest increases occurred in social-cultural awareness ($\Delta = 1.141$), educational value ($\Delta = 1.128$), and engagement ($\Delta = 1.123$).

The conventional-instruction group also showed pre-test to post-test increases across all dimensions ($p < .001$; Table 3), but the observed changes were smaller. Mean increases ranged from 0.360 to 0.443, with an average gain of 0.388, equivalent to approximately 12.9–15.7%. Improvement therefore occurred under both conditions, with consistently larger descriptive gains in the animation group.

Table 3
Pre–post comparisons within the experimental and control groups (N = 226 per group).

Variable	Experimental Pre-test Mean $\pm$ SD	Experimental Post-test Mean $\pm$ SD	Δ	Control Pre-test Mean $\pm$ SD	Control Post-test Mean $\pm$ SD	Δ	Experimental p	Control p
Interest	2.84 $\pm$ 0.40	3.92 $\pm$ 0.43	1.08	2.80 $\pm$ 0.45	3.16 $\pm$ 0.36	0.36	< .001	< .001
Comprehension	2.88 $\pm$ 0.40	3.89 $\pm$ 0.38	1.01	2.82 $\pm$ 0.37	3.18 $\pm$ 0.43	0.36	< .001	< .001
Knowledge acquisition	2.79 $\pm$ 0.38	3.90 $\pm$ 0.41	1.11	2.78 $\pm$ 0.42	3.18 $\pm$ 0.40	0.40	< .001	< .001
Emotional response	2.84 $\pm$ 0.41	3.92 $\pm$ 0.40	1.08	2.78 $\pm$ 0.36	3.15 $\pm$ 0.38	0.37	< .001	< .001
Educational value	2.78 $\pm$ 0.42	3.91 $\pm$ 0.42	1.13	2.81 $\pm$ 0.41	3.26 $\pm$ 0.41	0.44	< .001	< .001
Engagement	2.77 $\pm$ 0.39	3.90 $\pm$ 0.39	1.12	2.82 $\pm$ 0.40	3.19 $\pm$ 0.42	0.37	< .001	< .001
Social and cultural awareness	2.80 $\pm$ 0.41	3.94 $\pm$ 0.43	1.14	2.79 $\pm$ 0.35	3.21 $\pm$ 0.41	0.42	< .001	< .001
Heritage stewardship intention	2.81 $\pm$ 0.43	3.88 $\pm$ 0.40	1.07	2.79 $\pm$ 0.44	3.17 $\pm$ 0.42	0.38	< .001	< .001

Note. Values are presented as mean $\pm$ standard deviation; Δ is the mean post-test minus pre-test change. Figure 3 complements these descriptive values with confidence intervals and standardised between-group differences in change.

4.2.2. Knowledge acquisition and maintenance from T0 to T3

Before instruction, overall knowledge was low: the mean score was 2.15 of 7 (SD = 1.18), equivalent to 30.8% accuracy. Experimental and control means were 2.23 (31.9%) and 2.08 (29.6%), respectively; the unadjusted baseline difference was small and not statistically significant (difference = 0.16, 95% CI – 0.06 to 0.38; Hedges' $g = 0.14$; $p = .151$). The industrial-heritage-listing item was most difficult (13.5% correct overall), whereas the early-historical-development item had the highest baseline accuracy (44.2%). Baseline score and item-level distributions are reported in Table 4.

Table 4
Pre-test Overall Data Distribution (Dagushan knowledge test, n = 452).

No.	Knowledge Point	Experimental Group Accuracy	Control Group Accuracy	Overall Accuracy
1	Early historical development of Dagushan	47.3%	41.2%	44.2%
2	Japanese-era administration of Dagushan (1916)	34.1%	31.0%	32.5%
3	Soviet-era administration of Dagushan (1949)	30.5%	27.4%	29.0%
4	Dagushan's transformation of Ansteel (1956)	31.9%	25.2%	28.5%
5	Mineral resources of Dagushan (iron mine)	37.6%	38.5%	38.1%
6	Dagushan iron mine pit depth (-414 m)	31.9%	27.4%	29.6%
7	Industrial heritage listing (2017)	10.2%	16.8%	13.5%

At T1, overall mean accuracy increased to 64.7%. The experimental group averaged 5.20 correct answers (74.3%), compared with 3.85 (55.0%) in the control group. Experimental-group accuracy exceeded control-group accuracy on all seven items, with the largest descriptive gaps on historically specific and factual items. The industrial-heritage-listing item remained the most difficult, but T1 accuracy reached 52.7% in the experimental group versus 31.4% in the control group. Table 5 summarises the corresponding T1 totals and item-level accuracy for both instructional conditions.

Table 5
Post-test Overall Data Distribution (Dagushan knowledge test, n = 452).

No.	Knowledge Point	Experimental Group Accuracy	Control Group Accuracy	Overall Accuracy
1	Early historical development of Dagushan	86.3%	79.6%	83.0%
2	Japanese-era administration of Dagushan (1916)	73.9%	51.3%	62.6%
3	Soviet-era administration of Dagushan (1949)	75.2%	57.5%	66.4%
4	Dagushan's transformation of Ansteel (1956)	70.8%	47.8%	59.3%
5	Mineral resources of Dagushan (iron mine)	80.1%	57.1%	68.6%
6	Dagushan iron mine pit depth (-414 m)	81.4%	60.2%	70.8%
7	Industrial heritage listing (2017)	52.7%	31.4%	42.0%

Knowledge scores improved in both groups, with a larger individual-level gain in the experimental group. In the baseline-adjusted T1 model, the experimental group scored 1.37 correct answers higher than the control group (95%

CI 1.15 to 1.58). The mean T0–T1 gain was 2.97 answers in the experimental group and 1.77 in the control group; the difference in gain was 1.19 (95% CI 0.88 to 1.51; Hedges’ g = 0.69; Holm-adjusted p < .001). Item-level gains were positive overall but heterogeneous, so Fig. 4 distinguishes the total adjusted effect from the item-specific pattern.

Note

Panel (a) shows violin plots of pre- to post-test change in correct-answer count for the comparison and animation-supported conditions; the embedded box plot shows the median (centre line), IQR (box), 1.5 x IQR whiskers, and group mean (dot). Panel (b) shows the animation-supported minus comparison difference in post-test total score adjusted for baseline total and grade, with an HC3 95% confidence interval. Panel (c) shows the animation-supported condition’s advantage in pre-test to post-test item-accuracy change, calculated as the difference between conditions in percentage-point change.

The participant-linked longitudinal data showed attenuation after the immediate post-test in both conditions. Experimental-group means were 5.20 at T1, 4.90 at T2, and 4.51 at T3; control-group means were 3.85, 3.60, and 3.33. The condition-by-time interaction was significant, F (3, 1350) = 39.46, p < .001, indicating different score trajectories. Baseline-adjusted experimental–control differences were 1.37 at T1 (95% CI 1.15 to 1.58), 1.32 at T2 (95% CI 1.09 to 1.55), and 1.20 at T3 (95% CI 0.97 to 1.44), all Holm-adjusted p < .001. Hedges’ g declined from 1.15 at T1 to 1.06 at T2 and 0.93 at T3. Within the experimental group, mean change was – 0.30 from T1 to T2 and – 0.39 from T2 to T3; corresponding control-group changes were – 0.25 and – 0.27. All planned within-group declines remained significant after Holm adjustment. Thus, knowledge-test performance declined over time, but the animation group retained a substantial comparative advantage through T3. Table 6 presents the four-occasion means and unadjusted between-condition differences.

Table 6
Longitudinal knowledge-test performance and unadjusted between-group differences (n = 452).

Occasion	Experimental Mean ± SD	Control Mean ± SD	Difference (95% CI)	Hedges’ g
T0: pre-test	2.23 ± 1.21	2.08 ± 1.14	0.16 (– 0.06, 0.38)	0.14
T1: immediate	5.20 ± 1.16	3.85 ± 1.19	1.35 (1.14, 1.57)	1.15
T2: two weeks	4.90 ± 1.23	3.60 ± 1.24	1.31 (1.08, 1.53)	1.06
T3: 1–3 months	4.51 ± 1.30	3.33 ± 1.23	1.19 (0.95, 1.42)	0.93

4.3. Interpretive and implementation evidence

4.3.1. Student drawings

The strict archive audit retained 111 drawings with unambiguous grade provenance: 23 from Year 1, 17 from Year 2, 16 from Year 3, 18 from Year 4, 17 from Year 5, and 20 from Year 6. Four image identities carrying conflicting grade labels were excluded, and one within-grade duplicate was counted once. All remaining drawings were included irrespective of artistic quality or correspondence with the anticipated animation narrative. Figure 5 and Table 7 report the five-pass machine-assisted median consensus, retained as a fully grade-blinded sensitivity analysis. The subsequent paragraph reports the matched two-coder human verification.

Table 7

Exploratory grade-specific drawing indicators from five-pass machine-assisted median consensus (n = 111).

Consensus criterion	Year 1 (n = 23)	Year 2 (n = 17)	Year 3 (n = 16)	Year 4 (n = 18)	Year 5 (n = 17)	Year 6 (n = 20)
Narrative completeness	6 (26.1%)	11 (64.7%)	15 (93.8%)	14 (77.8%)	11 (64.7%)	14 (70.0%)
Explicit temporal sequencing	4 (17.4%)	10 (58.8%)	15 (93.8%)	11 (61.1%)	7 (41.2%)	11 (55.0%)
Explicit causal relation	4 (17.4%)	12 (70.6%)	15 (93.8%)	14 (77.8%)	11 (64.7%)	14 (70.0%)
Relational spatial understanding	8 (34.8%)	15 (88.2%)	16 (100.0%)	18 (100.0%)	16 (94.1%)	20 (100.0%)
Schematic/abstract representation	7 (30.4%)	11 (64.7%)	15 (93.8%)	16 (88.9%)	13 (76.5%)	15 (75.0%)
Value/affective expression	10 (43.5%)	13 (76.5%)	15 (93.8%)	16 (88.9%)	12 (70.6%)	19 (95.0%)

The revised results did not reproduce the simple lower–middle–upper gradient reported in the earlier restricted analysis. Year 1 contained the largest proportion of sparse or ambiguous drawings: 26.1% met the narrative-completeness threshold, 17.4% showed explicit temporal sequence, and 17.4% showed an explicit causal relation. Twelve of the 23 Year 1 drawings were flagged as uncertain by at least three coding passes, reflecting limited visible evidence rather than presumed lack of understanding. When those drawings were omitted in a sensitivity analysis, the corresponding Year 1 proportions rose to 54.5%, 36.4%, and 36.4%, respectively, but remained below the Year 3–4 pattern.

Performance increased sharply in Years 2 and 3. In Year 2, 64.7% of drawings met the narrative threshold, 58.8% showed explicit sequence, and 70.6% showed an explicit causal relation. Year 3 produced the most consistently integrated representations: 93.8% met each of the narrative, temporal, and causal thresholds; all drawings showed relational spatial understanding; and 93.8% contained schematic/abstract and value-oriented representation. Year 4 remained strong in narrative completeness (77.8%), causal representation (77.8%), spatial understanding (100%), abstraction (88.9%), and value expression (88.9%), although explicit temporal sequencing was lower (61.1%).

Years 5 and 6 retained high spatial representation (94.1% and 100%, respectively) but did not exceed Year 3 across all dimensions. Narrative completeness was 64.7% in Year 5 and 70.0% in Year 6; explicit temporal sequencing was 41.2% and 55.0%; and explicit causal relations were present in 64.7% and 70.0%. Value/affective expression was particularly frequent in Year 6 (95.0%). The nine-dimension total-score median was 3.0 in Year 1, 18.0 in Year 2, 20.0 in Year 3, 17.5 in Year 4, 13.0 in Year 5, and 18.0 in Year 6. Thus, higher grades were positively associated with total score overall (Spearman $\rho = .399$, Holm-adjusted $p < .001$), but the grade-specific pattern was non-monotonic.

Kruskal–Wallis tests indicated grade heterogeneity for all nine ordinal dimensions and the total score (all Holm-adjusted $p \leq .002$). These tests show that the grade distributions differed, not that maturation alone caused the differences. Unequal grade-specific task conditions, differences in drawing elaboration, and the concentration of uncertain drawings in Year 1 remain plausible explanations alongside developmental variation.

The five-pass machine-assisted analysis was stable after aggregation but less stable at the level of an individual pass. ICC(A,1) ranged from .351 to .621 and ICC(A,5) from .730 to .891. Exact five-way agreement was low, so these

values describe repeatability of the machine-assisted procedure rather than human reliability or criterion accuracy. After replacing HER with the three-person median, the final human total remained strongly aligned with the machine-assisted total (Spearman $\rho = .848$; mean absolute difference = 2.25 points). For HER specifically, machine–human correspondence was more modest ($\rho = .487$; mean absolute difference = 0.37), reinforcing the decision to treat heritage-meaning claims cautiously.

Human verification produced good reliability for the overall drawing score. Across the 111 matched drawings, the initial two full coders yielded $ICC(A,1) = .815$ for a single coder and $ICC(A,2) = .898$ for their mean. Dimension-specific $ICC(A,1)$ values were .839 for narrative completeness, .840 for temporal sequence, .633 for causal understanding, .684 for spatial understanding, .928 for schematic representation, .712 for value expression, .850 for environmental-systems understanding, and .769 for stewardship action. The initial HER value was lower (.302), and the first coder scored the 0–25 total an average of 3.55 points higher than the second, prompting the prespecified construct-level review rather than the recruitment of additional full-corpus coders.

The third coder therefore reviewed HER for all 120 archived files; reliability analysis used the same 111 grade-attributable drawings as the manuscript. The third coder HER mean was 0.77, between the first coder (1.05) and second coder (0.15). Three-rater HER reliability was $ICC(A,1) = .411$ for one coder and $ICC(A,3) = .677$ for the aggregate. Exact three-way agreement was 9.0%, but 96.4% of drawings were scored within one scale point across all three coders; only four drawings had a two-point range. The three-person median produced a mean HER score of 0.73 and was retained as the adjudicated value. HER interpretation therefore became more transparent and stable after aggregation but remained exploratory rather than a high-reliability construct.

Figure 6. Representative student reconstructions across Years 1–6. Three drawings per year group are shown to illustrate observable narrative, spatial, environmental, and heritage-related features in the coded corpus. Drawing identifiers are de-identified. The examples are illustrative and were not selected based on artistic quality; the complete corpus ($n = 111$), rather than the displayed subset, was used for all drawing analyses.

4.3.2. Teacher and heritage-guide accounts

The post-view analysis included seven primary teachers who participated in both interview rounds and four heritage guides. Four additional teachers contributed only to the pre-design round. The original grounded-theory analysis generated 38 open codes, eight axial dimensions, and three selective categories. For this paper, the appendix excerpts were reorganised by evidence type before interpretation. Direct observations were given greater evidential weight than reported student speech or prospective judgments; recommendations were analysed as implementation diagnoses rather than outcomes. This distinction matters because the appendix contains teacher observation, teacher interpretation, and nested student quotations within the same coding table.

The strongest teacher evidence concerned directly observed accessibility and classroom response. A04 reported that students “maintained interest while watching and did not get distracted,” an observation relevant to pacing and attention. A11 referred to students’ drawings and discussions when judging historical credibility and separately stated that knowledge tests and work showed improved historical-cultural understanding. A05 observed vivid characters in the drawings, although the inference that this demonstrated successful character design remained the teacher’s interpretation. These observations support usability and salience under the observed conditions; they do not independently estimate learning effects because no systematic observation schedule or blinded classroom observer was documented.

A second evidence layer comprised teacher-reported student speech. A02 recalled students identifying iron ore, the steel industry, and the mine’s century-long history. A01 reported the statement, “We should protect the environment

and not allow the mine to be further damaged,” while A10 reported sadness about damage and a wish to protect “our home.” A07 reported statements connecting litter to water pollution and rational mining to environmental protection and development. These excerpts provide qualitative support for historical and social–ecological meaning making, but they are nested quotations selected and relayed by teachers rather than direct student interview data. They therefore corroborate, but cannot replace, the student measures and drawing evidence.

Teachers’ critical and prospective comments exposed limits that the positive outcome themes alone would conceal. A06 relayed that some students wanted more three-dimensional visuals; A08 recommended differentiated background music to support emotional interpretation; and A06 proposed parent–child discussion of family industrial-heritage knowledge. In the formal improvement questions, interactive elements were mentioned six times. A10 recommended embedded knowledge quizzes, A09 suggested interactive narrative, A05 proposed humorous characters, A04 recommended field visits where feasible, and A11 proposed post-view puzzles or knowledge-card games. These are recommendations for the next program iteration; the records do not establish that they were implemented before the summative evaluation. Heritage guides supplied a different form of evidence—professional judgments about historical and technical credibility—and two explicitly acknowledged that they had not directly observed students. Their claims were therefore used for content validation, not learning-effect inference.

Mention-frequency analysis revealed an uneven outcome profile rather than uniformly strong effects. In the combined teacher–guide coding, stimulating interest in industrial heritage was mentioned 17 times; transmission of industrial-heritage history and emotional resonance were each mentioned 12 times; and storyline attractiveness was mentioned 11 times. Comprehensibility was mentioned 9 times, historical-information mastery and visual experience 8 times each, and sustainable-development concepts, learning interest, and social responsibility 7 times each. By contrast, heritage-protection awareness and enthusiasm for active learning were each mentioned twice, while advocacy of environmental action and environmental-action capability were each mentioned once. These are counts of coded mentions, not participant-level prevalence. The imbalance suggests that the animation was perceived primarily as an access, interest, and meaning-making resource, with substantially weaker evidence for participation or action.

Trustworthiness checks supported the stability of the broad interview framework but did not eliminate interpretive limitations. Coding was conducted by the primary researcher and independently reviewed by a second researcher; most NVivo nodes had Cohen’s $\kappa > .61$ and more than 75% exceeded .75. A supplementary Transformer-embedding/MLP text-classification model showed 87% overall agreement with the manual categories, with lower performance for semantically overlapping competency categories. This value is classifier–manual concordance, not proof that the coding or interview claims were accurate. Theoretical saturation was reported when new material fitted the three established categories and no new higher-order theme emerged. However, the retained appendix excerpts were predominantly supportive; apart from design critiques and recommendations, the record contains limited fully developed deviant cases. The analysis therefore reports both low-salience themes and improvement requests rather than claiming comprehensive negative-case saturation.

Taken together, the interview findings serve three bounded program-evaluation functions. First, pre-design teacher accounts explain why the resource was short, story-based, locally connected, and visually accessible. Second, post-view direct observations and professional judgments support classroom usability and historical credibility, while teacher-reported student speech offers secondary corroboration of historical and environmental meaning. Third, low-frequency action themes and explicit recommendations identify the next development cycle: greater interactivity, richer detail, humour, field connection, thematic games, differentiated sound and visuals, and family participation. The interviews do not provide a separate intervention-effect estimate. They show where the program appeared workable, which mechanisms participants considered plausible, and which components remained underdeveloped.

The source-specific findings were integrated only after the quantitative, drawing, and interview analyses were complete. Table 8 distinguishes convergence, divergence, and evidential boundaries across knowledge tests, learning-experience ratings, drawings, and teacher–guide accounts. These sources were treated as complementary evidence rather than as repeated validation of one effect. Four cross-source findings emerged.

Table 8
Joint display integrating formative interpretation and summative outcomes.

Evaluation claim	Student drawings	Teacher / guide evidence	Quantitative evidence	Integrated inference and boundary
Historical and narrative accessibility	Across 111 drawings, Year 1 provided the sparsest and most uncertain evidence; Years 3–4 showed the strongest integration, while Years 5–6 remained strong spatially without a uniform linear increase.	Direct observations supported attention and use of drawings/tests; teacher-reported student speech described historical facts. Guides provided professional credibility judgments.	Adjusted knowledge differences favoured animation at T1 (1.37), T2 (1.32), and T3 (1.20 correct answers); the condition-by-time interaction was significant.	Strong evidence of immediate access and partial maintenance of the core knowledge advantage; less evidence for transfer or uniform mastery of every detail.
Engagement and emotional response	Value/affective expression varied non-linearly and was frequent in Year 6; drawing evidence cannot independently establish engagement during viewing.	A04 directly observed sustained attention; emotional statements were teacher-reported student speech. Interest (17 mentions) and emotional resonance (12) were high-salience themes.	Adjusted differences favoured animation for emotional response (0.77), interest (0.76), and engagement (0.71), all Holm-adjusted $p < .001$.	Convergent short-term evidence; the design cannot separate effects of animation, narrative, and teacher facilitation.
Environmental and social–ecological meaning	Environmental and causal scores differed by grade, with the strongest integration in Year 3 and substantial within-grade variation elsewhere.	Teachers relayed student statements about damage, restoration, pollution, rational mining, and responsibility; these were nested reports, not direct student interviews.	Adjusted differences favoured animation for social-cultural awareness (0.73) and perceived knowledge acquisition (0.72).	Evidence supports emerging interpretation of a social–ecological transformation, but the knowledge test underrepresented systems reasoning.
Stewardship and action	Action orientation was coded separately from mining activity and general environmental imagery; single-pass agreement was only moderate.	Environmental action and action capability were each mentioned once; teachers recommended quizzes, games, field visits, and family discussion to move beyond intention.	Heritage stewardship intention favoured animation by 0.71 points; no behavioural or delayed outcome was measured.	Evidence supports disposition to care, not action competence or sustained stewardship.
Transferability and implementation	Non-linear and uncertainty-sensitive patterns indicate the need for differentiated prompts and verbal elicitation alongside drawings.	Phase 1 teachers emphasised short age-matched stories and local relevance; Phase 2 recommendations identified interactivity, detail, humour, field connection, and games.	Not directly tested.	The transferable unit is the place-based co-design and evaluation process, not the Dagushan content unchanged

4.4. Integrated findings across evidence sources

First, the strongest convergence concerned access rather than transformation. The animation group reported greater interest, emotional response, engagement, and social-cultural awareness, while teachers directly observed sustained attention and repeatedly described the story as accessible. Knowledge performance showed the same immediate advantage and retained a smaller advantage through T3. Together, these findings support the program as a means of making industrial heritage visible, memorable, and emotionally salient under the observed classroom conditions.

Second, structured performance exceeded open-ended reconstruction. The animation group answered more factual questions correctly at every post-instruction occasion, yet the drawings did not show uniformly strong reconstruction across grades. Year 1 evidence was frequently sparse or ambiguous, Years 3–4 showed the most integrated narrative, causal, spatial, and schematic representations, and Years 5–6 showed selective strengths rather than a monotonic increase. Correct answers on a repeated seven-item test therefore cannot be treated as equivalent to independently organising industrial, historical, and environmental relationships in an open task.

Third, caring was more evident than acting. Heritage stewardship intention favoured the animation condition, and teachers relayed statements about protecting the environment and using resources responsibly. However, heritage-protection awareness and active learning were each mentioned only twice in the combined interview coding, while environmental-action advocacy and capability were each mentioned once. Action orientation was also harder to distinguish reliably from general environmental imagery in drawings. The evidence therefore supports an emerging disposition to care, but not action competence, participation, or sustained stewardship behaviour.

Fourth, the program was usable but incomplete as an environmental education intervention. Teachers valued its short duration, local story, and visual accessibility, yet their recommendations consistently added elements absent from the evaluated program: embedded questions, interactive narrative, games, field visits, family discussion, and richer visual and auditory differentiation. These requests identify a program-development gap rather than a contradiction. The animation appeared mature as an introductory resource, but not as a self-sufficient pathway from awareness to inquiry and action.

5. Discussion

The central finding is not uniform effectiveness but an uneven distribution of educational outcomes. The animation-supported condition produced the clearest advantages in access, immediate experience, factual remembering, and partial retention over two months. Evidence of independent social-ecological reconstruction was less consistent, while direct evidence of action competence was scarce. This pattern exposes a central challenge for education in the Anthropocene: making extractive histories accessible and memorable is easier than enabling learners to interrogate their causes, evaluate competing development choices, or participate in alternative futures. Because Anthropocene education entails ethical deliberation and the development of individual and collective agency, access to information cannot itself be treated as evidence of responsible action (Olvitt 2017). Environmental learning outcomes may therefore be related without being interchangeable, and a favourable result in one domain cannot stand as evidence for transformation in another.

5.1. From accessibility to environmental meaning-making

Animation appears to have improved entry into a complex industrial history. All eight experience indices favoured the animation-supported condition, teachers reported stronger attention and comprehensibility, and the knowledge advantage remained at T3. These findings are consistent with multimedia accounts of how coordinated words and images can reduce the effort required to follow temporal and spatial change (Mayer 2024). They do not, however, demonstrate that learners independently reorganised those changes into social-ecological explanations. Research

showing that animation does not consistently outperform less dynamic materials similarly caution against treating perceptual fluency as conceptual understanding (Schubertová et al. 2023; Tani, Manuguerra, and Khan 2022).

The contrast between tests and drawings makes this boundary visible. Animation-supported students answered more repeated factual questions correctly, but their drawings varied in temporal order, causal connection, industrial detail, environmental consequence, and restoration. Years 3–4 showed the most integrated representations, while older students did not display uniformly greater sophistication. Drawing skill and task interpretation may partly explain that variation, so the drawings do not invalidate the test results. Instead, they show that remembering specified information and reconstructing relationships are different achievements. Accessibility is educationally valuable, but environmental meaning-making requires learners to select, connect, and explain relationships that were not already organised for them.

5.2. Non-substitutable outcomes in environmental education

The most important cross-source result is the separation between caring and acting. Questionnaire scores indicated stronger emotional response, social-cultural awareness, and stewardship intention, and teachers reported expressions of concern. Yet action advocacy and action capability were each coded only once in the interview corpus, while drawings rarely distinguished specific stewardship practices from generic green imagery. Because code frequency does not indicate prevalence across participants, this finding is interpreted only as limited qualitative evidence of action-oriented learning. Action competence requires knowledge of alternatives, reasoned judgement, confidence, participation, and persistence, none of which can be inferred from intention alone (Sass et al. 2020; Koskela and Paloniemi 2023; Oinonen et al. 2024). Related research in environmental art shows that memory and emotion can contribute to later action, but its direct measurement of behaviour also reinforces the need not to infer action from engagement alone (Kaufmann et al. 2023).

This distinction complicates evaluation models that allow positive knowledge or attitude indicators to represent programme success as a whole. Longitudinal research suggests that action competence develops unevenly and may not follow short-term attitudinal change (Olsson, Gericke, and Boeve-de Pauw 2022). The present evidence extends that caution to animation-supported industrial heritage education. Remembering, reconstructing, caring, and acting should be treated as analytically connected but non-substitutable outcomes. Their divergence is not a weakness to be averaged away. It identifies where an introductory resource contributes and where additional pedagogy is required.

5.3. Narrative coherence and the limits of critical participation

Narrative coherence is both a strength and a constraint. The three-generation story connected technological change, labour, family memory, environmental damage, and restoration in a form accessible to children. Story structure can support continuity and emotional orientation (Robin 2008), but a finished narrative may also supply key relationships and a preferred resolution in advance. Learners may therefore remember a coherent account without necessarily examining whose interests it represents, which experiences it omits, or how alternative interpretations might be constructed. This differs from participatory storying, where learners contribute perspectives, negotiate meanings, and exercise greater authorship (Häggström 2022; Kumpulainen et al. 2023).

For industrial heritage to function as a critical social-ecological text, learners must move beyond receiving a restoration narrative. They need opportunities to ask who benefited from extraction, who bore its environmental and social costs, how community dependence shaped available choices, and what collective responses remain possible. The teachers' requests for questions, games, field visits, family discussion, and richer interaction are therefore theoretically important rather than merely practical suggestions. They point to a transition from narrative reception to

investigation, contestation, and participation, consistent with critical and community-grounded place and heritage pedagogies in other settings (Häggström and Schmidt 2020; Riveiro-Rodríguez et al. 2021; Li and Shein 2023).

5.4. Implications for transferable programme design

The observed differences cannot be attributed to animation alone. Learners encountered a short visual narrative embedded in teacher-led instruction, and visualisation, story continuity, local memory, teacher clarification, novelty, and repeated testing may all have contributed. Both groups declined after T1, the adjusted difference narrowed from 1.37 to 1.20 correct answers by the fixed two-month follow-up, and the same seven items were reused. The defensible claim therefore concerns this instructional arrangement and maintenance of familiar test performance. It does not establish a universal animation effect, immunity to forgetting, or transfer to unfamiliar environmental problems.

What can transfer is a design and evaluation principle rather than the Dagushan story itself. Programmes can begin with a locally significant landscape, involve educators, heritage professionals, community knowledge holders, and affected groups, and translate historical and ecological relationships into age-appropriate media. They should then add structured opportunities for questioning, comparison, learner authorship, deliberation, and feasible action. Evaluation should document implementation and examine cognitive, emotional, interpretive, and participatory outcomes separately, as recommended by programme-evaluation research that links outcomes to programme characteristics and context (Stern, Powell, and Hill 2014; Garip et al. 2021; Gu et al. 2026). This approach preserves the accessibility of animation without mistaking accessibility for the completion of environmental education.

6. Limitations and future research

Several limitations bound the interpretation of the findings. The evaluation involved one school, and random sampling within grades did not constitute documented random assignment to instructional conditions. Unmeasured differences between conditions therefore cannot be excluded. T3 used the same seven items at a fixed two-month interval, while interim exposure was not monitored; the longitudinal results consequently indicate maintenance on familiar items rather than transfer. Implementation evidence was also limited because interviews were short, student statements were relayed through teachers, and no blind classroom observation was conducted. Drawings captured visible representation rather than the full extent of students' understanding. Grade labels were visible during manual coding, Year 1 ambiguity may partly reflect drawing production, and reliability for the industrial-heritage dimension remained limited ($ICC[A,1] = .411$). Future studies should include multiple schools, more transparent assignment procedures, alternate-form and longer-term assessments, implementation monitoring, grade-blinded drawing coding with brief student elicitation, and participatory tasks assessing deliberation, transfer, and action competence.

7. Conclusion

This mixed-methods evaluation indicates that *Centennial Mine: Dagushan* provided an accessible entry point to industrial heritage learning. The animation-supported arrangement produced higher immediate knowledge and learning-experience outcomes and retained a comparative knowledge advantage at the fixed two-month follow-up. However, remembering, reconstructing, caring, and acting did not advance equally. Open-ended reconstruction varied, evidence of action-oriented learning remained limited, and the repeated-item test did not establish transfer.

The study therefore contributes more than a general claim that animation works. It shows that accessibility, factual retention, social-ecological reconstruction, emotional orientation, and action competence are connected but non-substitutable outcomes. Industrial heritage animation can open an accessible route into the extractive histories that shape contemporary environmental conditions, but memorable narratives require subsequent opportunities to

question development choices, deliberate about social and environmental trade-offs, and participate in possible futures. The programme is therefore better understood as the beginning of an educational sequence for the Anthropocene than as a complete sustainability intervention. Whether narrative access develops into critical understanding and action competence remains an empirical question for participatory and transfer-oriented evaluation across other educational and cultural contexts.

Declarations

Ethics approval

The research was conducted in accordance with the Declaration of Helsinki, the Victoria University of Wellington Human Ethics Policy, and applicable institutional guidelines. Ethical approval was granted by the Human Ethics Committee at Victoria University of Wellington (School of Architecture; reference number 0000030449). Data collection took place in China, while the approving committee was based in New Zealand.

Consent to participate

Written informed consent to participate was obtained from all adult participants and from the parents or legal guardians of all participating minors. Participation was voluntary, and participants were informed of the study purpose, confidentiality arrangements, and their right to withdraw.

Data Availability

The raw data supporting this study are available on Zenodo at

<https://zenodo.org/records/22707538>

Author Contribution

X.B. conceived and designed the study, developed the educational materials, conducted the investigation and data collection, analysed the data, prepared the figures, and drafted the manuscript. A.B. and B.M. supervised the research and critically revised the manuscript. All authors reviewed and approved the final manuscript.

Acknowledgement

We are grateful to the students, teachers, heritage guides, museum professionals, and experts who contributed their time and perspectives to the development and evaluation of this research.

References

1. Author(s) (2025a) Anonymized for double-anonymous review.
2. Author(s) (2025b) Anonymized for double-anonymous review.
3. Author(s) (2026a) Anonymized for double-anonymous review.
4. Author(s) (2026b) Anonymized for double-anonymous review.
5. Cuenca-López JM, Martín-Cáceres MJ, Estepa-Giménez J (2021) Teacher training in heritage education: Good practices for citizenship education. *Humanit Social Sci Commun* 8(1):62. <https://doi.org/10.1057/s41599-021->

6. Garip G, Richardson M, Tinkler A, Glover S, Rees A (2021) Development and implementation of evaluation resources for a green outdoor educational program. *J Environ Educ* 52(1):25–39. <https://doi.org/10.1080/00958964.2020.1845588>
7. Garsdal J, Nordgren K (2026) Education and Bildung in the Anthropocene: Cartography of a field in tension. *Journal of Curriculum Studies*
8. <https://doi.org/10.1080/00220272.2026.2719513>
9. Gu X, Wang B, Xie K, Fu Y, Bexell SM (2026) The effectiveness and implementation of a humane education program in china: Promoting socio-emotional and ecological competencies among elementary school students. *J Environ Educ* 1–21. <https://doi.org/10.1080/00958964.2026.2651775>
10. Häggström M, Schmidt C (2020) Enhancing children's literacy and ecological literacy through critical place-based pedagogy. *Environ Educ Res* 26(12):1729–1745. <https://doi.org/10.1080/13504622.2020.1812537>
11. Häggström M (2022) Utilizing a storyline approach to facilitating pupils' agency in primary school sustainability education context. *J Environ Educ* 53(3):154–169. <https://doi.org/10.1080/00958964.2022.2067110>
12. Kaufmann DB, Palawat K, Sandhaus S, Buxner S, McMahon E, Ramírez-Andreotta M (2023) Communicating environmental data through art: The role of emotion and memory in evoking environmental action. *Humanit Social Sci Commun* 10:940. <https://doi.org/10.1057/s41599-023-02459-3>
13. Koskela I-M, Paloniemi R (2023) Learning and agency for sustainability transformations: Building on Bandura's theory of human agency. *Environ Educ Res* 29(1):164–178. <https://doi.org/10.1080/13504622.2022.2102153>
14. Kumpulainen K, Wong CC, Byman J, Renlund J, Vadeboncoeur JA (2023) Fostering children's ecological imagination with augmented storytelling. *J Environ Educ* 54(1):33–45. <https://doi.org/10.1080/00958964.2022.2152407>
15. Lambert J (2013) *Digital storytelling: Capturing lives, creating community*, 4th edn. Routledge, New York, NY, United States
16. Li J, Krishnamurthy S, Roders AP, Van Wesemael P (2020) Community participation in cultural heritage management: A systematic literature review comparing chinese and international practices. *Cities* 96:102476. <https://doi.org/10.1016/j.cities.2019.102476>
17. Li W-T, Shein PP (2023) Developing sense of place through a place-based Indigenous education for sustainable development curriculum. *Environ Educ Res* 29(5):692–714. <https://doi.org/10.1080/13504622.2022.2098933>
18. Lowan-Trudeau G (2023) Digital technologies and environmental education. *J Environ Educ* 54(1):1–7. <https://doi.org/10.1080/00958964.2022.2152413>
19. Mayer RE (2024) The past, present, and future of the cognitive theory of multimedia learning. *Educational Psychol Rev* 36(1):8. <https://doi.org/10.1007/s10648-023-09842-1>
20. Mendoza KR, Glaser N, Krüger JM, Yang M, Moeller K (2025) A systematic literature review on the impact of xr on pro-environmental knowledge, attitudes, and behaviors. *J Environ Educ* 56(2):85–125. <https://doi.org/10.1080/00958964.2025.2471299>
21. Morgado EMG, Licursi MB, Da Silva LLF (2024) Education and heritage: Memory, identity and educational dynamics. *Revista De Gestão Social E Ambiental* 18(8):e05960. <https://doi.org/10.24857/rgsa.v18n8-038>
22. Oinonen I, Seppälä T, Paloniemi R (2024) How does action competence explain young people's sustainability action? *Environ Educ Res* 30(4):499–518. <https://doi.org/10.1080/13504622.2023.2241675>
23. Olvitt LL (2017) Education in the Anthropocene: Ethico-moral dimensions and critical realist openings. *J Moral Educ* 46(4):396–409. <https://doi.org/10.1080/03057240.2017.1342613>

24. Olsson D, Gericke N, Boeve-de Pauw J (2022) The effectiveness of education for sustainable development revisited: A longitudinal study on secondary students' action competence for sustainability. *Environ Educ Res* 28(3):405–429. <https://doi.org/10.1080/13504622.2022.2033170>
25. Riveiro-Rodríguez T, Domínguez-Almansa A, López-Facal R, Izquierdo Rus T (2021) Place-Based Education and Heritage Education in in-service teacher training: Research on teaching practices in secondary schools in Galicia (NW Spain). *Humanit Social Sci Commun* 8:20. <https://doi.org/10.1057/s41599-020-00689-3>
26. Robin BR (2008) Digital storytelling: A powerful technology tool for the 21st century classroom. *Theory Into Pract* 47(3):220–228. <https://doi.org/10.1080/00405840802153916>
27. Russ A, Peters SJ, Krasny E M. and, Stedman RC (2015) Development of ecological place meaning in new york city. *J Environ Educ* 46(2):73–93. <https://doi.org/10.1080/00958964.2014.999743>
28. Sass W, Boeve-De Pauw J, Olsson D, Gericke N, De Maeyer S, Van Petegem P (2020) Redefining action competence: The case of sustainable development. *J Environ Educ* 51(4):292–305. <https://doi.org/10.1080/00958964.2020.1765132>
29. Sass W, Boeve-De Pauw J, Olsson D, Rieckmann M, Gericke N (2025) A quality education for sustainability teaching (quest) framework: Advancing action competence in sustainability education. *J Environ Educ* 56(6):417–434. <https://doi.org/10.1080/00958964.2025.2542937>
30. Schubertová K, Lukavský J, Drobná A, Volná K, Brom C (2023) Contextual animation in multimedia learning materials for pre-adolescents: The saga of null results continues. *Learn Instruction* 87:101803. <https://doi.org/10.1016/j.learninstruc.2023.101803>
31. Stern MJ, Powell RB, Hill D (2014) Environmental education program evaluation in the new millennium: What do we measure and what have we learned? *Environ Educ Res* 20(5):581–611. <https://doi.org/10.1080/13504622.2013.838749>
32. Tani M, Manuguerra M, Khan S (2022) Can videos affect learning outcomes? Evidence from an actual learning environment. *Education Tech Research Dev* 70(5):1675–1693. <https://doi.org/10.1007/s11423-022-10147-3>
33. Tugurian LP, Carrier SJ (2017) Children's environmental identity and the elementary science classroom. *J Environ Educ* 48(3):143–153. <https://doi.org/10.1080/00958964.2016.1191415>
34. Tzima S, Styliaras G, Bassounas A, Tzima M (2020) Harnessing the potential of storytelling and mobile technology in intangible cultural heritage: A case study in early childhood education in sustainability. *Sustainability* 12(22):9416. <https://doi.org/10.3390/su12229416>
35. Xie X, Wang C, Dai T, Chen Y (2026) An innovative model of digital pedagogy: Digital experience scenes design for cultural heritage properties based on multi-form spatial theory. *Humanit Social Sci Commun* 13:416
36. <https://doi.org/10.1057/s41599-026-06768-1>
37. UNESCO (2020) *Education for sustainable development: A roadmap*. Report for UNESCO (Paris, France). <https://www.unesco.org/en/articles/education-sustainable-development-roadmap>

Figures

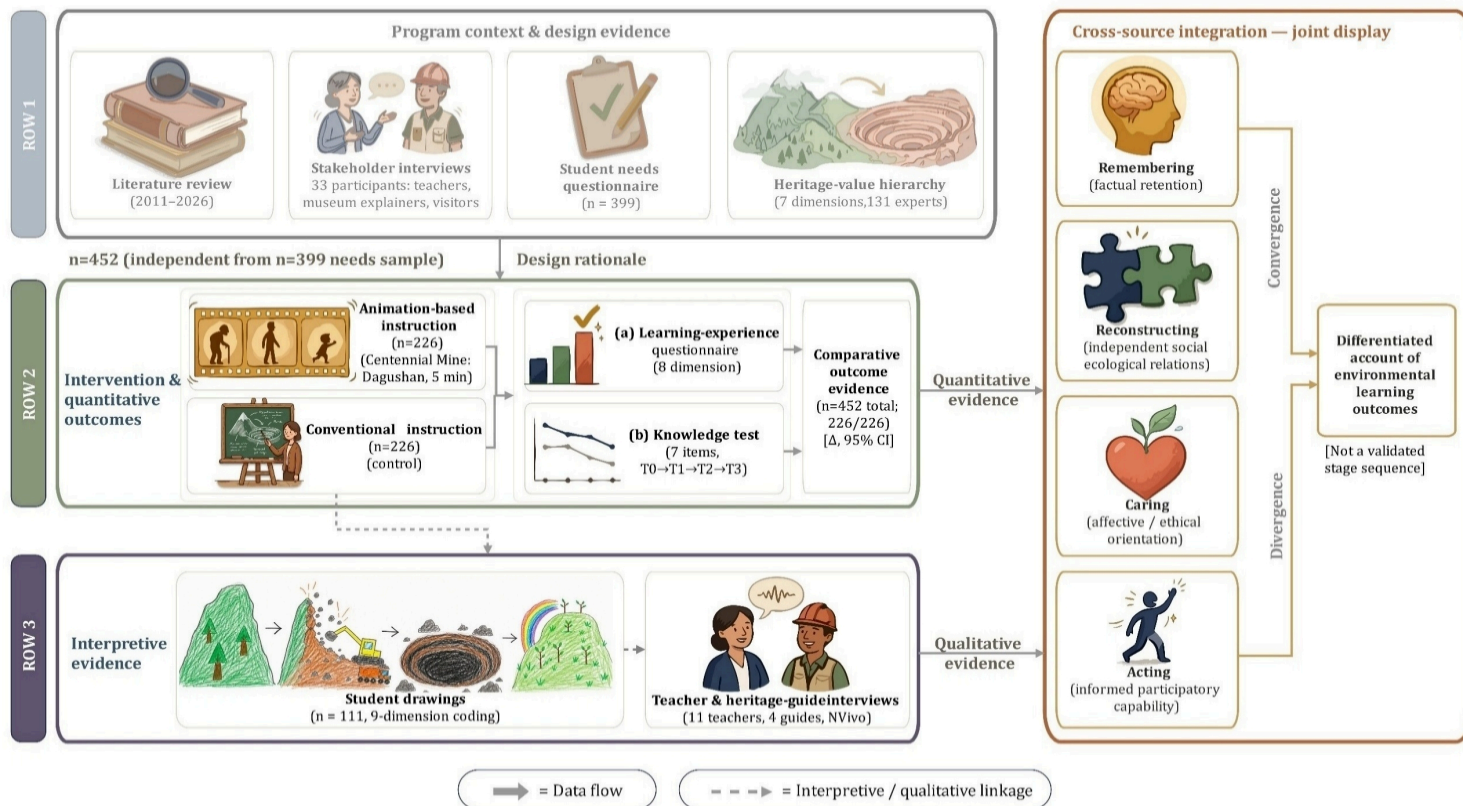


Figure 1

Evidence architecture of the outcome-focused mixed-methods evaluation. Row 1 summarises prior studies used only as background and design provenance; Rows 2-3 contain evidence analysed in this article. Cross-source integration distinguishes remembering, reconstructing, caring, and acting without treating them as chronological stages.



Figure 2

Baseline- and grade-adjusted post-test differences across eight learning-experience outcomes. Outcomes are ranked by the adjusted experimental–control difference; colour intensity represents the magnitude of the estimate, and the corresponding HC3 95% confidence interval is reported alongside it (n = 226 per group). All Holm-adjusted p values were < 0.001.

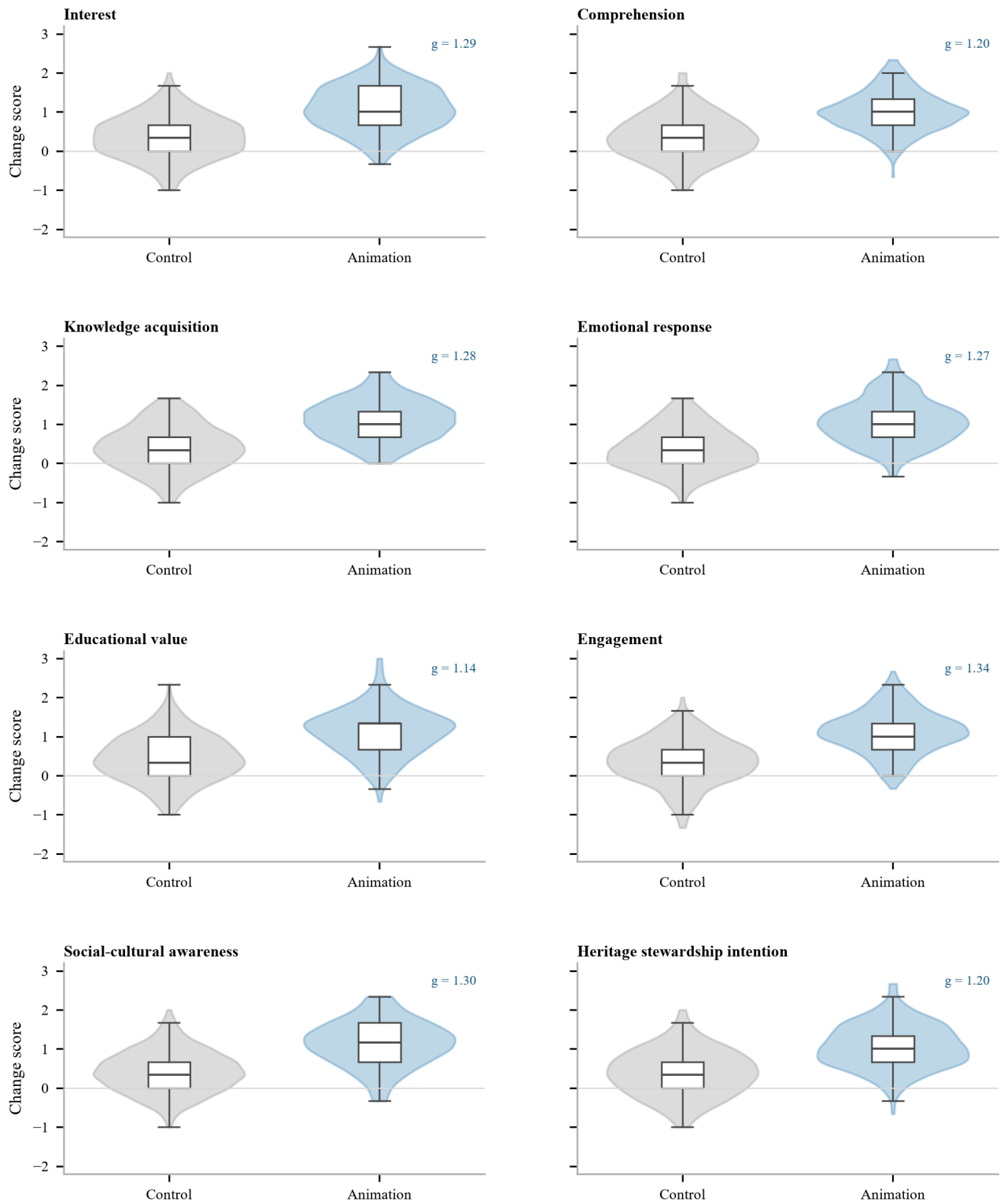


Figure 3

Distributions of pre-test to post-test change across eight learning-experience outcomes. Violin width represents the density of individual change scores; embedded box plots show the median and interquartile range. Grey denotes the control group and blue the animation group ($n = 226$ per group). Each panel reports Hedges' g for the between-group difference in change; bootstrap confidence intervals are reported in the source data.

Note. Violin plots show the distributions of individual pre–post change scores (post-test minus pre-test) for the experimental (blue) and control (grey) groups. Embedded boxplots indicate the median and interquartile range (IQR), and black dots represent the mean change scores. Positive values indicate improvements from pre-test to post-test. Panels represent (a) composite learning experience, (b) interest, (c) comprehension, (d) knowledge acquisition, (e) emotional response, (f) educational value, (g) engagement, (h) social and cultural awareness, and (i) heritage stewardship intention.

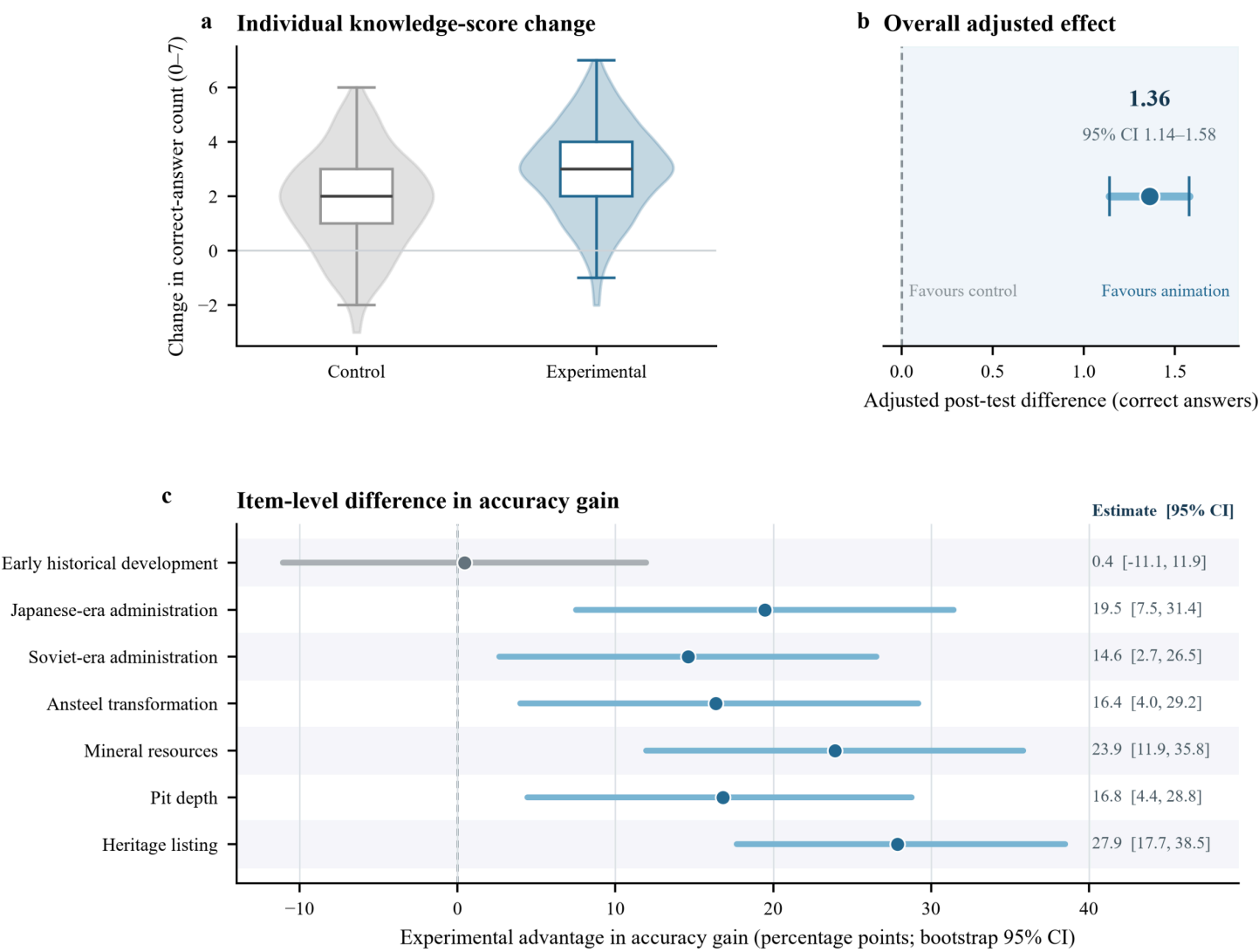


Figure 4

Knowledge-test outcomes (n = 452). (a) Distribution of individual change in total correct answers, shown with violin and box plots. (b) Experimental–control post-test difference adjusted for baseline total and grade, with HC3 95% confidence interval. (c) Item-level experimental advantage in pre-test to post-test accuracy change, with percentile-bootstrap 95% confidence intervals.

Note. Panel (a) shows violin plots of pre- to post-test change in correct-answer count for the comparison and animation-supported conditions; the embedded box plot shows the median (centre line), IQR (box), 1.5 x IQR whiskers, and group mean (dot). Panel (b) shows the animation-supported minus comparison difference in post-test total score adjusted for baseline total and grade, with an HC3 95% confidence interval. Panel (c) shows the animation-supported

condition's advantage in pre-test to post-test item-accuracy change, calculated as the difference between conditions in percentage-point change.

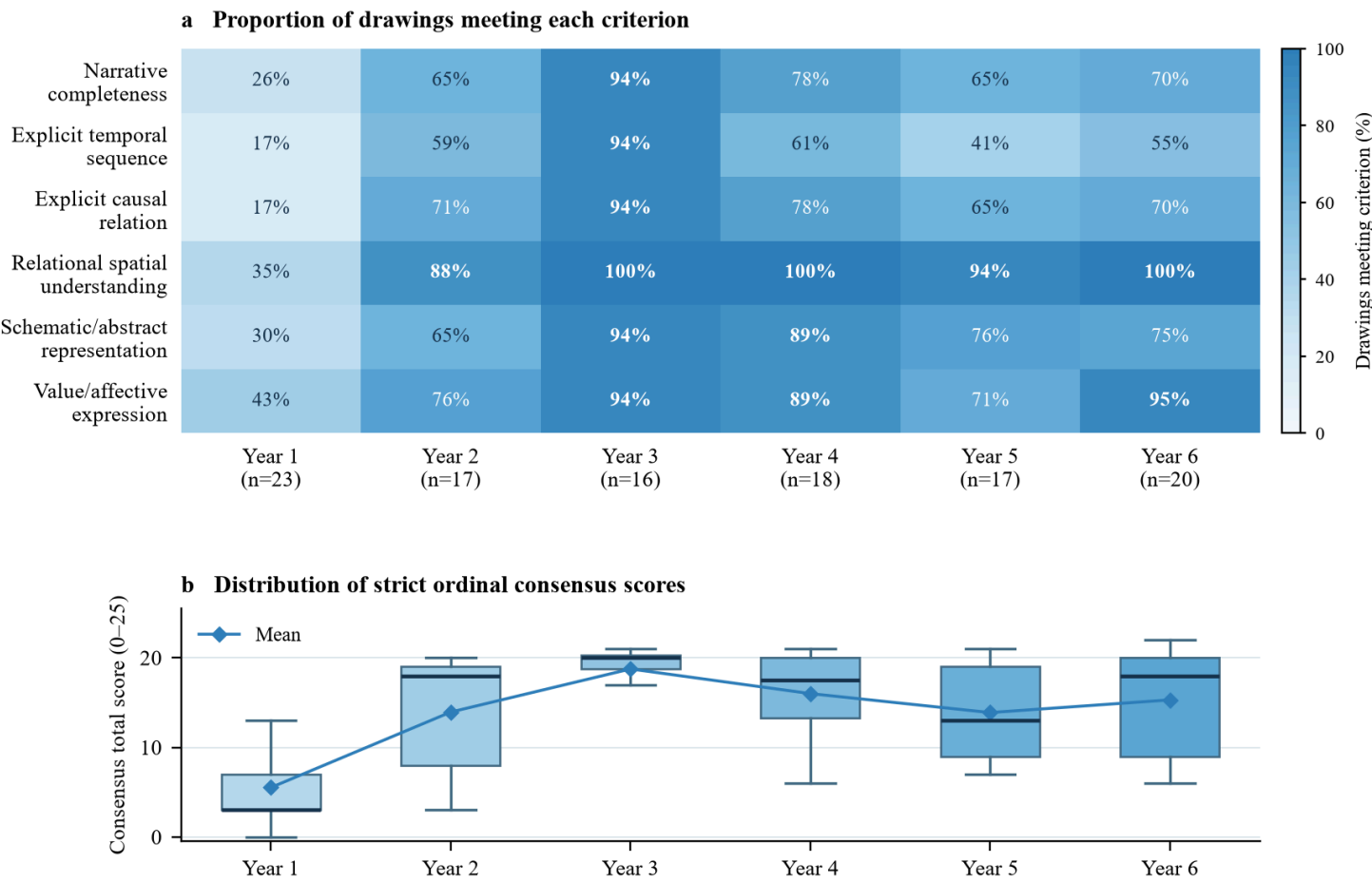


Figure 5

Exploratory machine-assisted analysis of drawing-based narrative reconstruction by grade (n = 111). (a) Percentages of drawings meeting six manuscript-compatible consensus criteria. (b) Distributions of the nine-dimension consensus total score (range 0-25); boxes show the median and interquartile range, whiskers extend to 1.5 x IQR, and diamonds show means. Scores are medians of five grade-blinded machine-assisted coding passes. The figure is a sensitivity analysis and does not replace the human coding results or represent individual longitudinal development.

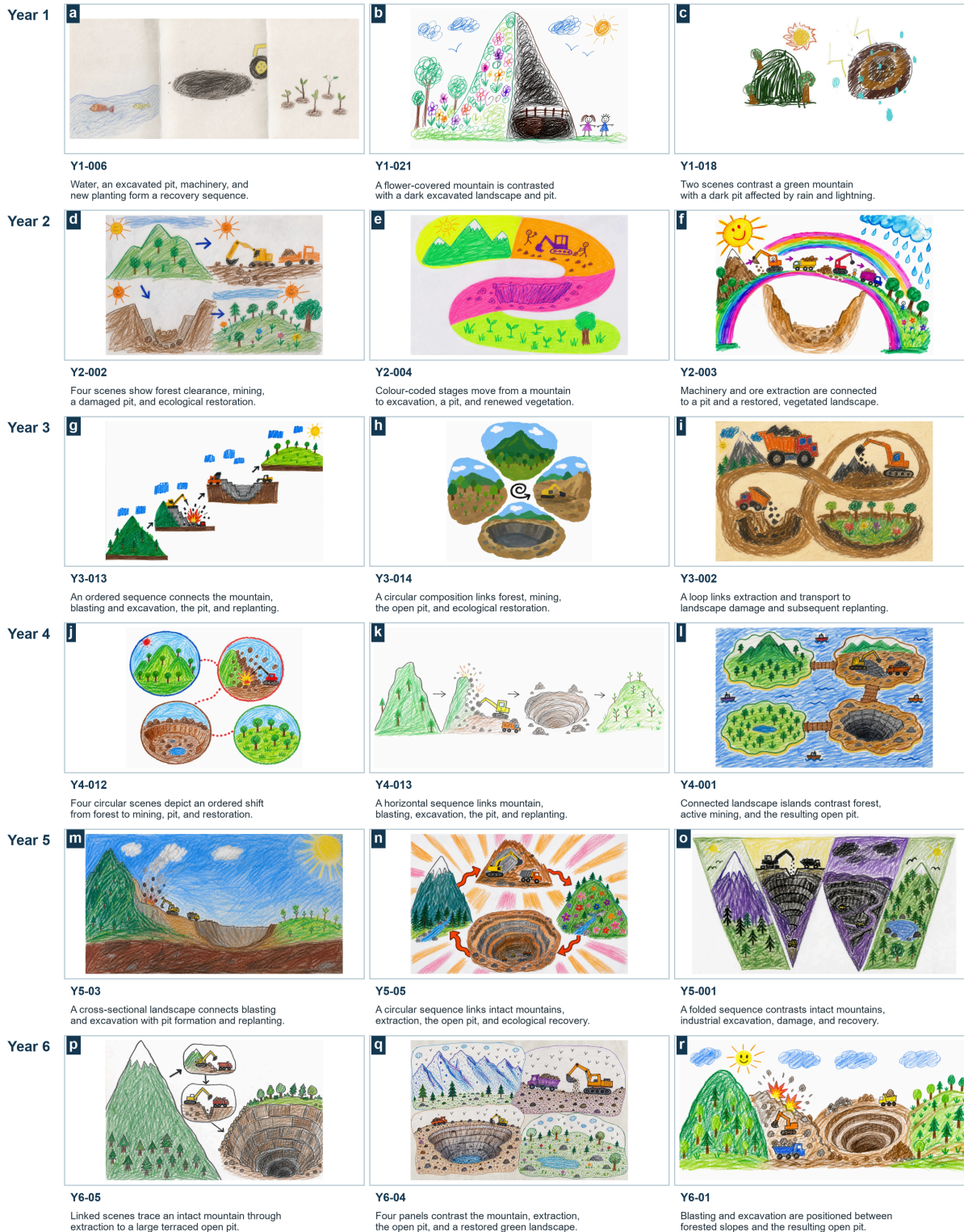


Figure 6

Representative student reconstructions across Years 1–6. Three drawings per year group are shown to illustrate observable narrative, spatial, environmental, and heritage-related features in the coded corpus. Drawing identifiers are de-identified. The examples are illustrative and were not selected based on artistic quality; the complete corpus (n = 111), rather than the displayed subset, was used for all drawing analyses.