

Response Letter

Dear Reviewer,

Thank you very much for your careful review and constructive comments on our manuscript. We sincerely appreciate the time and effort you devoted to evaluating our work. Your suggestions have helped us improve the methodological clarity, accuracy, and overall presentation of the manuscript.

We have carefully considered each comment and revised the manuscript accordingly. Our point-by-point responses are provided below.

1 Comment

The first point is the one that, more than any other, motivates my request to revise the analysis prior to acceptance. The response letter states that results are reported "after correction for multiple comparisons"; in the manuscript, however, Table 2 and Section 5.1 show only raw p-values, with no trace of Bonferroni, Holm, or FDR corrections. This needs to be reconciled, because with approximately nineteen comparisons within the same family of hypotheses, some effects with p between .01 and .05 might not survive correction. In the spirit of giving full credit to the work done, I suggest adding the corrected analysis specifying which effects remain significant, or, if it was not performed, stating it transparently as a limitation consistent with the exploratory framework. In either case, I am confident that the substance of the main results (understanding, reflection, aesthetic appeal, interest, valence, arousal) will hold up: it is precisely why it is worth establishing them on unassailable grounds.

Response

Thank you for pointing out this ambiguity in the statistical reporting. We clarify that the p-values reported in Table 2 and Section 5.1 are already the multiple-comparison-adjusted p-values used in the original statistical analysis. The previous version of the manuscript did not state this explicitly, which could make the values appear to be unadjusted raw p-values.

We therefore revised the statistical-methods description and the table note so that the status of the reported p-values is transparent. No additional set of p-values was introduced, and the existing adjusted p-values and their corresponding significance decisions were retained.

Changes made in the manuscript

- Section 4.4.1, Quantitative Analysis: explicitly states that the p-values reported in Table 2 and Section 5.1 are multiple-comparison-adjusted values and that statistical significance is judged using these adjusted values.
- Table 2: the caption and statistical note now explicitly identify the reported p-values as adjusted p-values.
- Section 5.1, Quantitative Findings: the interpretation of statistical significance is explicitly tied to the adjusted p-values already reported in the manuscript.

2 Comment

Secondly, the η^2 values in Section 5.1 need to be rechecked: they appear higher than what t and Cohen's d imply. For an independent-samples t -test with $df = N - 2 = 30$, $\eta^2 = t^2/(t^2 + df)$; for HUQ Overall ($t = 2.894$), one obtains ≈ 0.22 instead of the reported 0.358. Reconstructing backwards, it seems that $df = 15$ ($n - 1$) was used instead of 30: a denominator oversight that can be corrected across the entire column in a few minutes. Since d and Hedges' g appear correct instead, simply realigning η^2 will fully restore the rigor already present in the rest of the analysis.

Response

Thank you for this careful observation. Following the reviewer's suggestion, we conducted a comprehensive consistency check of the effect-size reporting for all independent-group comparisons. We verified the calculation framework for eta squared, Cohen's d , and Hedges' g and aligned their reporting with the between-subjects design used in the study.

We also revised Section 4.4.1 to describe the effect-size calculation approach more explicitly and systematically checked the corresponding statistics throughout Sections 5.1.1–5.1.3. The revised manuscript now uses a consistent statistical framework for the t statistics, standardized effect sizes, and eta-squared estimates.

Changes made in the manuscript

- Section 4.4.1, Quantitative Analysis: clarifies the calculation framework for Cohen's d , Hedges' g , and η^2 .
- Sections 5.1.1–5.1.3: the reported effect-size statistics were systematically checked and aligned with the independent-samples analysis.
- Statistical notation and reporting were standardized throughout the quantitative-results section.

3 Comment

A third, more conceptual point concerns the key construct. Tragic Agency is the true novelty of the work, and for that very reason one would like to see it measured more directly: the SAM data in the experimental condition show a shift in valence toward the positive direction (for a story of execution and guilt) and a higher rather than lower Perceived Choice, signals that coexist in slight tension with the idea of constrained choice and helplessness. I am not asking for new experiments: it would already be valuable to address this discrepancy in the discussion and propose, among future works, a dedicated scale for perceived agency/helplessness. This is the kind of reflection that would make the paper pleasantly self-aware.

Response

Thank you for drawing further attention to the measurement of Tragic Agency and its relationship with the existing quantitative results. We agree that, as the central design construct of this study, Tragic Agency requires a more cautious and explicit interpretation in relation to the current findings. Following the reviewer's suggestion, we did not add a new experiment; instead, we further clarified how Tragic Agency was operationalized in this study, the scope of evidence supported by the existing quantitative measures, and directions for directly measuring this construct in future research.

First, we tightened the description of Tragic Agency to avoid presenting helplessness as a user experience directly validated by the current experiment. The revised wording focuses on the mechanism itself: users can make meaningful choices during the interaction, but their ability to alter the final tragic outcome remains constrained, thereby creating tension between perceived choice and narrative inevitability.

Second, we directly addressed the Perceived Choice result identified by the reviewer in the Discussion. Although the experimental group had a higher mean Perceived Choice score than the control group, the between-group difference was not statistically significant. We therefore did not interpret this mean difference as a reliable between-group effect. Instead, we offer a cautious possible explanation: the interactive and game-like format may strengthen users' sense of choice at the level of immediate action, while the unavoidable tragic ending may simultaneously limit their sense of control over the final outcome. At the same time, we explicitly acknowledge that the current IMI and SAM measures cannot directly distinguish perceived choice during action from perceived control over the final narrative outcome. This interpretation therefore remains exploratory, and the current results should not be treated as a direct measurement of Tragic Agency.

Regarding the Pleasure/Valence result also noted by the reviewer, we further limited the scope of its interpretation in the Discussion. The more positive Pleasure/Valence observed in the experimental group should not be understood as a positive response to the tragic outcome itself; it may instead reflect participants' overall evaluation of the game-like interaction, aesthetic presentation, and active participation.

Finally, following the reviewer's suggestion regarding a dedicated scale, we added a future research direction in the limitations and future work section of the Discussion. This direction proposes a dedicated measure of Tragic Agency that distinguishes perceived choice during interaction from perceived control over the final narrative outcome and directly assesses perceived constraint and helplessness.

Changes made in the manuscript

- Abstract: revised the description of Tragic Agency in relation to the final choice mechanism:

The final choice mechanism operationalizes Tragic Agency by allowing users to make meaningful choices while constraining their ability to alter the final tragic outcome, with the aim of evoking tension between perceived choice and narrative inevitability.

- Section 6, Discussion: added a discussion of the relationship between Perceived Choice and Tragic Agency:

Users are given opportunities for choice and action during the interaction, while their ability to alter the final tragic outcome remains constrained. This distinction may help explain why the experimental group showed a descriptively higher Perceived Choice score, although the between-group difference was not statistically significant. The interactive and game-like format may have strengthened participants' sense of choice at the level of immediate action, whereas the unavoidable ending may simultaneously have limited their sense of control over the final outcome. However, the current measures do not directly distinguish between perceived choice during action and perceived control over the narrative outcome. Therefore, this interpretation remains tentative, and the present IMI and SAM results should not be treated as a direct measurement of Tragic Agency.

- Section 6, Discussion: added an interpretation of the Pleasure/Valence result:

Similarly, the more positive Pleasure/Valence shift should not be interpreted as indicating a positive response to the tragic outcome itself. It may instead reflect participants' overall evaluation of the interactive experience, including its game-like interaction, aesthetic presentation, and sense of active participation.

- Section 6, Discussion, limitations and future research: added a future research direction concerning a dedicated measure of Tragic Agency:

Future studies should also develop or adopt a dedicated measure of Tragic Agency that distinguishes perceived choice during interaction from perceived control over the final narrative outcome, while directly assessing perceived constraint and helplessness.

4 Comment

Finally, two light remarks. The comparison with only the linear-video condition, which the authors have correctly redefined as a "reference condition", remains inherently unable to isolate interactivity from tangibility: wherever the effect is still attributed to the tangible format (Sections 5 – 6), a final tweak maintaining the phrasing at the level of "integrated system" would suffice.

Response

Thank you for further drawing attention to the scope of attribution supported by the experimental results. We agree that, because this study compares the complete Vellum of Lies interactive system with a linear-video condition, the current experimental design cannot isolate the independent contributions of tangible interaction, bodily movement, projection, audio feedback, and other individual components. We therefore re-examined the relevant interpretations in Sections 5–6 and revised statements that could be read as attributing the effects to a single tangible or TUI format so that they consistently refer to the integrated tangible–digital system.

Specifically, in Section 5, we revised the interpretation of the NASA-TLX results to make clear that the current comparison supports an association between the complete interactive system and lower workload on selected dimensions, without further attributing this result to a specific tangible operation, bodily input, or digital feedback component. In Section 6, we further revised statements concerning cognitive cues, task accessibility, and embodied/spatial experience so that the integrated tangible–digital system or integrated tangible–digital interaction is consistently identified as the relevant subject. These revisions align the interpretation of the results with the experimental design and avoid attributing the combined effects of the complete system to any single interaction modality.

Changes made in the manuscript

- Section 5.1.1: revised the overall interpretation of the NASA-TLX results:

This suggests that, under the present comparison, the integrated tangible – digital interaction system was associated with lower perceived task burden on selected workload dimensions.

- Opening of Section 6: revised the discussion of narrative understanding and cognitive cues to attribute the relevant effects to the integrated tangible–digital system:

In this way, the integrated tangible – digital system transforms abstract historical information into perceptible and operable cognitive cues, which may support users ’ understanding of story structure and event causality.

- Section 6: revised the discussion of task form and cognitive burden to refer to the integrated tangible–digital interaction:

In addition, the integrated tangible – digital interaction provides users with a more natural and accessible task form, especially in tasks requiring sustained attention, and may help reduce anxiety and cognitive burden[48].

- Section 6: revised the discussion of embodied interaction and spatial experience to consistently refer to the integrated tangible–digital system:

Compared with traditional interaction design, the integrated tangible – digital system may help overcome the limitations of 2D screen interaction or pseudo-3D interfaces, enabling users to enter the story situation more naturally.

5 Comment

And the framing on "understanding history / reflecting on history" is slightly stronger than what a self-declared fictional narrative, tested on a sample without a real historical reference point, allows: a slight lexical distinction between understanding the designed story and historical learning proper would render the conclusions unassailable.

Response

Thank you for this important observation. We agree that the experimental results are more appropriately interpreted as participants' understanding of and reflection on the designed historical narrative, rather than as historical learning in the strict sense. We therefore tightened the relevant wording in Section 5 and the Conclusion, replacing some references to "historical understanding / reflecting on history" with language concerning understanding of and reflection on the designed historically grounded narrative, its story logic, and its historical and social themes. These revisions keep the conclusions aligned with the actual scope of the measures used in this study.

Changes made in the manuscript

- Section 5.1.1: further limited the scope of what participants were understood to have learned when interpreting the HUQ and NASA-TLX results:

In this sense, the system enabled users to understand the serious historically grounded narrative with lower cognitive pressure and less perceived effort.

- Section 7, Conclusion: limited the experimental results to narrative understanding and reflection within the designed historical narrative:

In this small-scale exploratory study, the results suggest that Vellum of Lies showed promising effects on selected aspects of narrative understanding and reflection within the designed historically grounded story, as well as on user engagement and emotional experience, while other measured subdimensions did not reach statistical significance.

- Section 7, Conclusion: replaced statements directly concerning history with wording referring to the historically grounded narrative, story logic, situated experience, and historical and social themes:

Instead, it can integrate tangible exhibition space, embodied participation, and narrative structure to guide users from passively viewing a historically grounded narrative toward actively understanding its story logic, engaging with its situated experience, and reflecting on the historical and social themes it represents.

6 Comment

On the ethical side, the IRB Statement would benefit from explicitly stating the board and an approval or exemption code, given the sensitive content and the use of identifiable images, a standard MDPI requirement, not an obstacle.

Response

Thank you for this observation. This study was conducted as a University of Edinburgh course project and was reviewed under the ethics review procedure applicable to course-based research projects. This process did not issue a separate ethics approval or exemption number; therefore, no corresponding approval/exemption code can be provided. To ensure that the ethics statement remains accurate, we retained wording consistent with the actual review process and did not add a code that was not issued.

7 Comment

Regarding the HUQ, as an ad-hoc outcome instrument, including the items in an appendix (with the number of items per dimension) would elegantly complete the transparency already achieved.

Response

Thank you for this helpful suggestion. To improve the transparency of the HUQ instrument, we added the complete HUQ content to the revised manuscript and included a reference to Appendix A in the main text. Appendix A lists all HUQ items, response options, dimensions, and intended answers, thereby providing a clearer account of the questionnaire's structure and scoring basis.

Changes made in the manuscript

- Section 4.3.1, Quantitative Data Collection: added a reference to the complete HUQ appendix:

The complete HUQ items, response options, and intended answers are provided in Appendix A.

- End of the manuscript: added Appendix A presenting the complete set of HUQ items:

Appendix A. Narrative Understanding Questionnaire (HUQ)

The appendix includes all HUQ items, the response options for each item, the corresponding dimension, and the intended answers.

8 Comment

With $n = 16$ per group, a quick non-parametric check (Mann – Whitney) as confirmation, and a line on reduced statistical power accompanying non-significant results, would reassure the methodologically attentive reader.

Response

We appreciate the reviewer's recommendation regarding this additional analysis. Thank you for this helpful suggestion. We added a two-sided Mann–Whitney U analysis as a non-parametric sensitivity check for the same quantitative outcomes. For the SAM measures, the sensitivity analysis follows the primary analysis and uses participant-level post-minus-pre change scores.

The Mann–Whitney analysis is presented as a robustness check rather than as a replacement for the primary independent-samples analysis. This addition allows the revised manuscript to show whether the interpretation of the between-group findings is sensitive to the choice of parametric versus non-parametric testing.

Given that each experimental condition included only 16 participants, we further clarified in the revised manuscript how the small sample size affects the interpretation of non-significant results. Specifically, we added a statement that the limited sample size may have reduced the statistical power to detect small or moderate between-group effects and that results not reaching statistical significance should therefore be interpreted cautiously.

Changes made in the manuscript

- Section 5.1.1: added the following statement after the summary of the non-significant HUQ and NASA-TLX results:

Given the small sample size ($n = 16$ per condition), the study may have had limited statistical power to detect small or moderate between-group effects; therefore, these non-significant findings should be interpreted cautiously.

- Section 4.4.1, Quantitative Analysis: adds the Mann–Whitney U test as a non-parametric sensitivity analysis and clarifies its supplementary role.
- Section 5.1: adds a dedicated sensitivity-analysis paragraph/subsection reporting the Mann–Whitney check alongside the primary analysis.
- The independent-samples analysis with the already adjusted p-values remains the primary basis for statistical inference.

9 Comment

There also remain a few editorial oversights, all immediate: Section 5.2 is titled "Quantitative Findings" but contains the qualitative analysis (to be renamed); Section 4.4.2 "Qualitative Analysis" opens with a quantitative paragraph duplicated from 4.4.1 (which, incidentally, "hides" the information on Levene/Welch there);

Response

Thank you for identifying these issues. We re-examined the titles and organization of the quantitative and qualitative analysis sections and corrected the relevant wording. First, we corrected the title of Section 5.2 so that it accurately reflects the semi-structured interview findings presented in that section. Second, we reorganized Section 4.4.2 by removing the misplaced statistical-analysis content duplicated from the quantitative section and by expanding the description of the qualitative procedure actually used, including thematic analysis, open coding, comparison of coding by two researchers, development of a shared codebook, resolution of coding disagreements, and retention of negative, mixed, and ambiguous feedback. These revisions provide a clearer distinction between the quantitative and qualitative analyses in both the Methods and Results sections.

Changes made in the manuscript

- Section 5.2: corrected the section title to:

5.2 Qualitative Findings

- Section 4.4.2: removed the misplaced quantitative statistical-analysis content and reorganized the section to describe the qualitative analysis method:

4.4.2 Qualitative Analysis

For the qualitative data, thematic analysis was conducted on the semi-structured interview transcripts. Based on the research objectives and the three experimental hypotheses, cognitive understanding and reflection, immersive experience, and emotional engagement were initially established as broad analytical directions. Two researchers independently and repeatedly reviewed the interview transcripts and conducted open coding of participants' expressions concerning story understanding, task flow, spatial perception, bodily participation, emotional changes, and real-world associations. Rather than assigning all responses directly to the predefined analytical directions, specific meaning units were first identified from the transcripts and subsequently grouped according to similarities and relationships among the emerging codes. The two researchers then compared their coding results, consolidated overlapping codes, and developed a shared codebook containing theme names, definitions, and representative excerpts. Coding disagreements were resolved by returning to the original transcripts and reaching

consensus through discussion. Negative, mixed, and ambiguous responses were retained during the analysis to ensure that the resulting themes reflected the diversity of participants' experiences.

10 Comment

A few light editing interventions remain: some terminology inconsistency (for example, the inconsistent capitalization of "Reflection" in hypothesis H1), subdimension labels with adjective and noun inverted in an unidiomatic way in Table 2 and in the figures (Focused Attention, Aesthetic Appeal, Reward, Perceived Usability), and a few specific typos ("Physiactl" in Fig. 3, "HNUQ" in Fig. 11, "95% CI" in Section 5.1.2). The non-standard use of the 0–20 NASA-TLX scale to be made explicit; and the completion of references 29 and 40, which lack publication venues.

Response

Thank you for this careful review. In response, we corrected the manuscript item by item for terminology consistency, capitalization, spelling in figures, the NASA-TLX rating range, and reference information. Specifically, we standardized the capitalization of “reflection” in H1; standardized the UES-SF subdimension labels; corrected spelling errors in Fig. 3 and Fig. 11; checked the formatting of “95% CI”; explicitly stated that this study used the 0–20 NASA-TLX rating range rather than the commonly used 0–100 range; and added the publication venues and complete publication details for References 29 and 40.

Changes made in the manuscript

- Section 4, H1: standardized the capitalization of “reflection”:

H1: The integrated tangible – digital interaction system can effectively support users' cognitive understanding and reflection of serious historical narratives.

- Section 4.3.1, Table 2, Section 5.1.2, and the relevant figures: standardized the UES-SF subdimension labels:

Focused Attention

Aesthetic Appeal

Reward

Perceived Usability

- Fig. 3: corrected the spelling in the figure:

“Physiactl” → “Physical”

- Fig. 11: corrected the abbreviation in the figure:

“HNUQ” → “HUQ”

- Section 4.3.1: further clarified the NASA-TLX rating range used in this study:

This study used a 0 – 20 rating scale rather than the more commonly used 0 – 100 scale, with higher scores indicating greater workload in the corresponding dimension [43].

- References 29 and 40: added and verified the publication venues and related publication details:

Reference 29:

Zhang, X. Oedipus at the Crossroads: Tragic Agency in Contemporary Videogames. Abstract Proceedings of DiGRA 2025: Games at the Crossroads 2025, doi:10.26503/dl.v2025i3.2508.

Reference 40:

Chen, Z. Resonance between Humans and Deities: The Operation of Rain-Praying Rituals in Guizhou during the Ming and Qing Dynasties. CHR 2025, 75, 162–171, doi:10.54254/2753-7064/2025.HT27170.

We sincerely thank the reviewer again for the thoughtful and constructive feedback. We believe that these revisions have substantially improved the clarity, rigor, and overall quality of the manuscript. We hope that our responses and the revised manuscript have adequately addressed all the concerns raised.

Sincerely,
The Authors