

Review of manuscript egusphere-2026-2372

“Gamified Disaster Education. A Systematic Review of Serious Games on Dealing with Natural Hazards and of their Impacts and Outcomes for Users”

General assessment

The topic addressed by this manuscript is both important and highly relevant to Geoscience Communication. Serious games are increasingly used in disaster-risk education and risk communication, yet the field still lacks a sufficiently clear and consolidated understanding of what these tools actually achieve, how their impacts are evaluated, and how strong the available evidence is. For this reason, I regard a systematic review of this literature as very much needed. Precisely because I think this review has the potential to become a useful reference for the field, some of the comments below are deliberately strict. They are intended to help the authors make the review methodologically stronger, more transparent, and more reliable as a basis for future research and practice.

The manuscript addresses a relevant question: what evidence currently exists regarding the impacts of serious games used for natural-hazard education and risk communication? The distinction between studies that simply describe games, those evaluating usability, and those attempting to evaluate impacts on players is useful. The review also brings together a relatively large body of literature (132 publications) and highlights several patterns that are potentially valuable for researchers and practitioners, particularly the predominance of digital games, the strong representation of flood and earthquake applications, and the concentration of evaluation studies on knowledge and self-efficacy.

I nevertheless have concerns about the methodological transparency of the review and, more importantly, about the strength of some conclusions drawn from the evidence collected. At present, the manuscript moves somewhat uneasily between a systematic mapping of the literature and an assessment of effectiveness. The former is reasonably convincing and potentially useful; the latter requires more caution. The 42 “effectiveness evaluation studies” are highly heterogeneous, most use relatively weak designs for causal inference, and the manuscript does not report a formal assessment of study quality or risk of bias. Yet several passages synthesise the evidence in terms that imply effectiveness, and the Discussion itself makes statements about the quality of the underlying studies without explaining how quality was systematically assessed.

There are also important reproducibility issues in the search, screening and coding procedures. These are particularly relevant because the manuscript explicitly argues for more structured, standardised and transparent evaluation research. The review itself should meet a similarly high standard of methodological transparency. I consider the paper potentially suitable for Geoscience Communication, and the topic is clearly within the journal’s remit. However, I think major revision is required before the review can support the claims currently made.

Major comments

1. The literature search needs to be updated and described in substantially more detail

The search was conducted on 16 January 2025, whereas the preprint entered discussion in May 2026. Given the sharp increase in publications documented by the authors themselves, this is a substantial gap for a review presented as an updated overview of the current research landscape. The manuscript should therefore be updated with a new search conducted as close as reasonably possible to submission/revision.

This is not simply a matter of adding a few recent references. A search update may alter the frequencies reported throughout the Results, the composition of the 42 effectiveness studies, and potentially some of the conclusions.

The search strategy also requires greater reproducibility. The manuscript states that Scopus, Web of Science and “EBSCOhost” were searched, but EBSCOhost is a platform hosting multiple databases. Please specify which EBSCO database(s) were used. Similarly, it is not clear which bibliographic fields were searched in each database (title, abstract, keywords, topic, etc.), or whether the displayed search string was used identically across all three platforms. Database-specific search syntax and field restrictions should be reported.

The sensitivity of the search strategy also deserves consideration. The eligibility criteria include studies described using terms such as “disaster games” or simply “games”, yet the game-related part of the search is relatively restricted. The authors should explain how they assessed whether this strategy was sufficiently sensitive to capture studies not explicitly using “serious game”, “game-based learning”, “gamification”, or “educational game” terminology.

I am not asking for an unnecessarily complex search strategy, but the manuscript needs enough information for another researcher to reproduce it and enough justification to demonstrate that the search is fit for the review’s stated scope.

2. Screening, extraction and coding procedures are currently insufficiently described

The manuscript states that the systematic review was conducted following the PRISMA 2020 statement. I suggest that the authors provide a completed PRISMA 2020 checklist, either as supplementary material or alongside the revision, indicating where each applicable item is addressed in the manuscript.

At present, I could not identify complete reporting for several core items, including the information sources and database interfaces (Item 6), database-specific search strategies (Item 7), selection process (Item 8), data-collection process (Item 9), some elements of data-item definition and synthesis methodology (Items 10 and 13), registration/protocol information (Item 24), and availability of extracted data and review materials (Item 27).

Not all PRISMA items will necessarily require an analysis—for example, meta-analysis-specific procedures are not relevant if no meta-analysis was undertaken—but the checklist would make it clear which items are applicable, which procedures were performed, and where they are reported. Where a procedure was not undertaken, the manuscript should state this rather than leaving the reader uncertain.

Section 2.3 states that “at least one of the authors” read and documented each publication. This leaves several important aspects of the systematic review process unclear.

I could not determine whether title/abstract screening and full-text screening were conducted independently by more than one reviewer; whether any subset of the records was double-screened; how disagreements about eligibility were resolved; whether data extraction or categorical coding was independently checked; whether a coding manual or operational definitions were established beforehand; or whether agreement between coders was assessed.

This matters particularly because several central classifications require researcher judgement. Examples include deciding whether something constitutes a “complete” serious game rather than gamification, whether a study qualifies as a usability or effectiveness evaluation, assigning games to DRM phases, classifying learning objectives, identifying theoretical frameworks, and grouping measured outcomes into knowledge, competencies, motivation/intention/behaviour, and psychological effects.

The manuscript does not necessarily need to have used two independent coders for every record, but it does need to explain the procedure actually followed and how coding reliability was protected. If no independent verification was performed, this should be transparently acknowledged as a limitation.

I would also encourage the authors to provide the coding framework or extraction sheet as supplementary or repository material. This would substantially improve the reproducibility of the review.

3. The manuscript needs to distinguish more clearly between mapping reported outcomes and demonstrating effectiveness

This is my main conceptual concern.

The review labels 42 papers as “effectiveness evaluation studies” and RQ3 asks which impacts and outcomes serious games “have” for players. However, the methodological summary shows that only eight studies included a control-group comparison, whereas most relied on pre-post designs. Follow-up assessment was rare. Some outcomes were based on self-reported knowledge or behavioural intentions rather than demonstrated knowledge or actual behaviour.

These studies can certainly provide useful evidence, but they do not all provide the same level of evidence for effectiveness.

The manuscript does not report a systematic risk-of-bias or methodological-quality assessment. Nevertheless, the Discussion states that evaluation quality varied considerably and sometimes describes approaches as close to “pseudo-evaluations”. Without predefined appraisal criteria, it is difficult for the reader to understand how these judgements were reached.

I see two reasonable routes here. One option would be to introduce an appropriate and transparent methodological-quality appraisal of the effectiveness studies. Given the heterogeneity of designs, this does not need to become a meta-analysis, nor do I think a meta-analysis is necessarily appropriate.

Alternatively, the authors could more explicitly frame the work as a systematic evidence map/descriptive synthesis of reported evaluation outcomes, and revise the language throughout accordingly. In that case, statements such as “playing a serious game can indeed improve the players’ knowledge” should be replaced by formulations that reflect what the reviewed studies actually demonstrate, for example that multiple studies reported pre-post increases in knowledge, while the strength of causal evidence varies substantially across designs.

This distinction is especially important for the conclusion that serious games have the potential to increase knowledge and self-efficacy. I think this conclusion is plausible from the reviewed literature, but the confidence that can be placed in it should be explicitly related to the quality and design of the underlying studies.

4. Operational definitions and outcome classifications need to be made more reproducible

The manuscript rightly points out that terminology is inconsistent across the field, but some of this ambiguity also enters the review’s own analytical framework.

For example, the authors state in Section 3.2.3 that some studies treat risk awareness, perception and understanding similarly to knowledge. This is an important observation, but it raises the question of how the reviewers themselves classified such outcomes.

The manuscript should explain the decision rules used to distinguish factual versus self-perceived knowledge; knowledge versus risk awareness; risk awareness versus risk perception; competencies/skills; motivation; behavioural intention; actual behaviour; and psychological outcomes.

This is particularly important because the frequency with which each outcome category appears is one of the main findings of the review.

A similar issue applies to the central distinction among no evaluation, usability evaluation, and effectiveness evaluation. The conceptual explanation given earlier in the manuscript is helpful, but the coding criteria used in practice should be stated more explicitly. What minimum evidence was necessary for a publication to be classified as an effectiveness evaluation? How were studies containing both usability and impact outcomes classified? Were categories mutually exclusive by design?

The manuscript would become considerably stronger if these classifications could be reproduced by another research group using the same definitions.

5. There are numerical inconsistencies in the core review results that need to be resolved

There is a clear inconsistency between Figure 1 and the Abstract/Results regarding the evaluation categories.

Figure 1 reports: No evaluation: $n = 23$; Usability evaluation: $n = 67$; Effectiveness evaluation: $n = 42$.

The Abstract and Section 3.1 report: No evaluation: $n = 17$; Usability evaluation: $n = 73$; Effectiveness evaluation: $n = 42$.

Both sets sum to 132, so this appears to be a classification/version inconsistency rather than a simple arithmetic error. Because this categorisation determines which publications enter the detailed effectiveness analysis, it should be corrected throughout and the underlying dataset checked.

There is a second issue in the description of study countries. For example, the manuscript reports USA $n = 20$ as 14.4% and Italy $n = 12$ as 8.6%. Those percentages are not percentages of the 132 publications. They appear instead to use a denominator of approximately 139 country occurrences, presumably because multi-country studies were counted more than once. If so, this is perfectly legitimate, but the denominator needs to be explicitly stated. At present, the text refers to percentages of “publications”, which makes the calculation difficult to interpret.

Please conduct a consistency check of all frequencies, percentages, figures and the Appendix after the database is finalised.

6. Several statements in the Discussion go beyond what can be established from the review

There are a number of places where interpretation should be separated more clearly from evidence.

Most importantly, the statement that “publication bias can be assumed” because negative and null findings were rarely reported is too strong. The observed pattern is certainly compatible with publication or selective-reporting bias, but this review has not formally assessed publication bias. I suggest stating that the scarcity of null findings raises the possibility of publication/selective-reporting bias.

Similarly, the manuscript states that none of the 132 studies “had been pre-registered”. If this means that none of the publications reported preregistration, please say exactly that. Absence of a preregistration statement in the article does not necessarily demonstrate that a study was not preregistered elsewhere.

The ethics discussion has the same problem. The manuscript correctly reports whether an ethics approval or ethics statement was mentioned, but subsequently states that the majority of EE studies “do not have ethics approvals”. From the information analysed, the defensible conclusion is that they did not report an ethics approval, unless the authors independently verified this elsewhere.

Finally, some explanations for temporal publication patterns are highly speculative. The proposed connections between the 2016 increase and the Sendai Framework, and between the post-2020 pattern and COVID-19/distanced learning, are possible interpretations but are not tested in this review. They should be explicitly presented as hypotheses rather than explanations of the observed pattern.

These changes would not weaken the paper; they would make the distinction between review evidence and reviewer interpretation much clearer.

7. The Data availability statement is not consistent with the study that is reported

The manuscript states: “No data sets were used in this article.”

However, Section 2.3 explains that information from the publications was systematically collected in SoSciSurvey and analysed in SPSS and Microsoft Excel. That extracted and coded information is itself a research dataset underpinning the numerical results of the systematic review.

I therefore do not think the current Data availability statement is appropriate.

At minimum, the authors should provide the extracted review dataset underlying the figures, frequencies and classifications, together with enough documentation to interpret its variables. Ideally, the coding framework and, where possible, the search/export information should also be archived.

This is particularly relevant because reproducibility and transparency are central to the manuscript’s own recommendations for the field.

Minor comments

1. Title: “Gamified Disaster Education” is broader than the actual scope and may also create conceptual confusion because the manuscript explicitly excludes interventions consisting only of gamification elements. A title referring directly to serious games for natural-hazard education/risk communication would more precisely reflect the review.
2. Abstract: Please correct the evaluation-category numbers once the discrepancy with Figure 1 has been resolved.
3. Introduction, Section 1.1: “Recent educational discourses, reinforced by the COVID-19 pandemic underscores...” should be grammatically corrected (“underscore”).
4. The phrase “a long examination of the topic in a motivating way” in Section 1.1 is difficult to understand and should be rephrased.
5. Section 1.3: “survey are commonly used” should be corrected.
6. The authors use both “impact evaluation” and “effectiveness evaluation” for the same subset of studies. I suggest selecting one term and using it consistently, particularly because “effectiveness” carries a relatively strong causal implication.
7. Figure 1: “Recorts” appears several times and should be corrected to “Records”.
8. Figure 2: because the literature search took place on 16 January 2025, the 2025 bar represents only a very small fraction of the year. This should be stated directly in the figure or caption; otherwise the apparent fall in 2025 can easily be misread as a publication trend.

9. Figure 3: the meaning of the colour intensity is not explained by a legend, and several country labels/numbers are difficult to read at manuscript scale. Please clarify the visual encoding and improve readability. The counting rule for multi-country studies should also be included in the caption.
10. Figure 4: “adressed” should be corrected to “addressed”. Since individual publications may address multiple hazards, the caption should explicitly state that categories are non-exclusive.
11. Table 1: theories, theoretical models, instructional-design frameworks and research approaches are combined in one list. Given that RQ1 specifically asks about behavioural theories or models, it may help the reader if these different types of frameworks are distinguished rather than presented as equivalent.
12. Section 3.2.2: I suggest revising “the majority of EE studies ... do not have ethics approvals” to “do not report ethics approval”, for the reason discussed above.
13. Section 3.2.3: “participants perceptions” requires a possessive apostrophe.
14. Discussion: “In a several theoretical approaches” should read “In several theoretical approaches”; “meta-analysis show” should be “meta-analyses show” or “a meta-analysis shows”.
15. Section 4.1: “Although ethics approvals required...” is missing “are”.
16. “metanalytical” should be corrected to “meta-analytical”.
17. Section 4.2: “games themself” should be “games themselves”.
18. Appendix Table A1: “Conclusive list” is unusual here; “Complete list” or “List of all publications included...” would be clearer.
19. A final language edit would be worthwhile. The English is generally understandable and the argument can be followed, but there are enough grammatical and idiomatic issues to distract occasionally from the scientific content.
20. The manuscript cites at line 365 “Gargiulo et al. (2024)” in the discussion of risk perception, although two Gargiulo et al. publications from 2024 are present in the reference list and elsewhere distinguished as 2024a and 2024b. The relevant citation should be disambiguated consistently.
21. “Feng et al. (2020, 2021)” appears (line 342) in the discussion of preparedness training although two Feng et al. publications from 2020 are listed as 2020a and 2020b. The relevant 2020 publication should be specified.
22. The Bareford et al. (2021) DOI is written as “https://doi.org/https://doi.org/...”, resulting in a duplicated DOI prefix.
23. The Witte (1992) reference contains the same duplicated “https://doi.org/https://doi.org/...” construction.
24. Bandura (1977a), “Self-efficacy: Toward a Unifying Theory of Behavioral Change”: the DOI currently given in the manuscript does not correspond to the Psychological Review citation listed. The DOI links to *Bandura, A. Self-efficacy: Toward a unifying theory of behavioral change, Advances in Behaviour Research and Therapy, Volume 1, Issue 4, Pages 139-161, ISSN 0146-6402, 1978* and not *Bandura, A.: Self-efficacy: Toward a Unifying Theory of Behavioral Change, Psychological Review, 84, 191–215, 1977*. The authors should verify which version they intend to cite and make the journal, year, pagination and DOI internally consistent.
25. The DOI formatting for Witte (1992) should be corrected by removing the duplicated DOI prefix.

Because the search predates the manuscript by more than one year, the reference set should be re-generated after an updated database search rather than supplemented manually. I did not identify evidence that would justify questioning the existence of the broader body of references. The main bibliographic concerns are consistency, DOI accuracy and the need to update the review period.

Recommendation

Minor revisions

The manuscript addresses a worthwhile question and contains a potentially valuable synthesis, but a revision is needed before the conclusions about serious-game effectiveness are sufficiently supported. The essential points are to update and fully document the literature search, clarify and strengthen the transparency of screening and coding procedures,

resolve the numerical inconsistencies, distinguish descriptive evidence on reported outcomes from stronger claims of effectiveness, and either formally appraise methodological quality or substantially temper conclusions according to the strength of the underlying study designs. The data generated through the systematic review should also be made available, or the authors should provide a clear justification if this is not possible. Once these issues are addressed, I think the manuscript could make a useful contribution to Geoscience Communication.