Reviews for jovi-2023-files-uncertainty-round1

Review #1

Completed: 11-11-2023 14:09

Recommendation: Revisions Required

Conflict Declaration

I declare that I have no known conflicts of interest with the authors.

Review

The research tests if practice with 2-D tasks helps people perform better when they make decisions based on more complex 3-D visualizations. The study had a three-part design. Participants were divided into three groups for training: one used ellipse displays, another used scatter displays on a 2-D plane, and the third group had no training. In the transfer phase, they worked with a 3-D snowy terrain model, using either ellipse or scatter displays. Participants were randomly placed into one of six groups based on these training and transfer conditions. Additionally, everyone faced two levels of task difficulty during both the training and transfer stages. The authors have reported that training with 2-D visualizations helped participants respond faster in a 3-D task, but it didn't make them more accurate. They also discussed what these findings mean for designing visual tools.

Overall, I believe that the issue studied in this article is important and cutting-edge. If solid research conclusions can be drawn, it would facilitate effective expression of complex data, especially in terms of uncertainty perception and analysis. However, there are several major issues in the article that need to be addressed.

First, the research motivation of this paper is not very clear. This work is a further expansion of the authors' previous research, which, according to the author's own words, aims "to see if performance boosts from practice transfer to new contexts as well as to a change in visualization." The authors define the change in context as a shift from a 2D to a 3D display scenario, without altering the visual encoding form of data uncertainty. My first question is, under what circumstances do we need to train in a 2D scenario but infer in a 3D scenario? With the decreasing cost of constructing 3D content, it stands to reason that training directly in a specific 3D scenario would be more effective than training in a 2D scenario.

Second, as the authors point out, completing tasks in a 3D scenario involves understanding the 3D terrain and manipulating the viewpoint in 3D space. Another question is whether the differences in the experimental results are due to the training received by the subjects, the 3D terrain itself, or differences caused by various observation perspectives. For example, the radii of an elliptical contour are related to the spread of data uncertainty, and a change in the viewpoint can distort this elliptical contour, affecting the user's judgment of the radii and, consequently, their judgment of uncertainty. For another example, if the viewpoint is moved to orthogonally above the 3D terrain, the change in the elliptical contour, compared to a 2D scenario, is entirely due to the undulations of the 3D terrain. This raises the question:

does the user's understanding of the uncertainty encoded by the contour get affected by the terrain's undulations? The experiments designed in this paper do not fully consider these factors, leading me to have doubts about the conclusions drawn from the experimental results.

Third, the analysis of the experimental data in the article is not thorough, and there are instances where the analysis results do not correspond with the data. Here are a few examples.

- In the analysis of Figure 4, the authors state, "Parameter estimates of the zero-inflated beta model (Figure 4) show that both and [the mean and sample size (precision) parameters, respectively, of the beta distribution] increased more slowly for untrained participants compared to participants trained with ellipse or scatter stimuli." However, looking at the first row of the four graphs on the left and middle of Figure 4, only two graphs show the blue curve increasing slower than the green and orange curves, while the data in the other two graphs do not support the authors' conclusion. Additionally, the authors claim that " and for untrained participants ended up higher by the end of the simple block compared to trained participants." While this is true in the simple-ellipse-mean-parameter and simple-ellipse-sample-size-parameter graphs where the blue curve eventually is higher than the others, the other two graphs do not show this trend. Therefore, it is inappropriate for the authors to conclude that "performance of the untrained participants ended up slightly better than that of the participants who received training."
- It can also be observed that people trained with 2-D ellipse stimuli performed better on the 3-D scatter task of simple difficulty compared to untrained participants or those trained with 2-D scatter stimuli. This interesting finding was not discussed by the authors.
- In analyzing Figure 5, the authors state, "Error distance analysis showed similar results to the response ratio analysis. Overall, performance (in terms of average distance between selected and correct response) was similar across training conditions (Figure 5)." However, upon closer examination, there are differences between the two. The authors should provide a more detailed comparison and description, as this single sentence analysis is insufficient.
- In the analysis of Figure 7, the authors state, "There was a mismatch between observed mean response times on the first two trials and the model's posterior predictive distribution over those two trials." However, this difference is not observable in the figure. In addition, the authors state, "Overall, response times in the complex block were similar to those in the simple block, suggesting that the added complexity did not substantially slow down task performance." But it is obvious that there are differences.

Finally, this paper has presentation issues and is difficult to follow, requiring further improvement. Here are some of my suggestions. - The hypotheses of the experiment should be clearly stated, along with what corresponds to each experimental phase. - The data analysis could be clarified by highlighting or enlarging relevant parts in the data figures. - Although this paper is an extension of the authors' previous work, the author should not assume that readers are already familiar with the relevant background and experimental

setup. Key concepts should be clearly explained and described. For example, what is meant by “best location”? What are the quantitative criteria for determining the best location? Why choose two different models to analyze accuracy? - Each equation should be numbered and the notations within the equation clearly explained. - The figure numbers are incorrect.

Based on the above points, my recommendation for this paper is a major revision.

Openness

The authors provided an OSF link in the paper, which includes relevant code and data. However, I was unable to find information regarding the questionnaire.

Classification

Empirical Research - Quantitative

Recommendation

Major Revisions

Revisions Requested

- Provide a detailed explanation of the research background and its significance.
- Modify the experimental design to either eliminate the factor of viewpoint or add experimental components to comprehensively analyze the impact of viewpoint changes on user decision-making.
- Conduct a more comprehensive and detailed analysis of the experimental data.
- Fix presentation issues.

Reviewer Name

anonymous

ORCID

N/A

Review #2

Completed: 12-11-2023 06:17

Recommendation: Revisions Required

Conflict Declaration

I declare that I have no known conflicts of interest with the authors.

Review

This paper presents a study on transfer learning from 2D to 3D environments for an uncertainty integration task. It provides evidence the complex interplay between 2D and 3D representations for similar or identical tasks, and contributes to our knowledge regarding the effect of task cueing or training on task completion using novel (to the user) representations. Interestingly, unpracticed participants often performed the tasks more accurately, but persisted in having longer response times with the novel representation. The experimental design is coherent, well thought-out, and the results are interesting. However, the paper does struggle to articulate its logic and content satisfactorily at points, such as the link between the hypotheses and the experimental design. Another weakness it the lack of consideration of existing work in 3D visualization and cartography that would provide additional relevance/areas of contribution for this research, and potentially provide more explanation for the study results. I believe this paper is ready for publication subject to minor revisions. In my opinion the revisions are minor since I see no fundamental issues within the study that need to be addressed.

Openness

The results/analysis and stimuli are accessible via OSF, but the study questionnaires need to be provided. I would prefer the stimuli within the study be provided with their own page rather than a generic home page containing links to some of the authors other projects and variants not used in this study. Demographic and exploratory analysis were provided, including helpful visual summaries and source code. However, the study questionnaires including the Graph Literacy Scale, the Berlin Numeracy Scale, Cognitive Reflection Task, Computer Self Efficacy Questionnaire, and Intrinsic Motivation Inventory were not included as far as I could tell within the OSF repositories. No additional supplementary material was provided.

Classification

Empirical Research - Quantitative

Recommendation

Minor Revisions

Revisions Requested

Note that this is a copy and paste from Word which alters the formatting; the uploaded reviewer response document should be referenced.

Requested Changes

1. 3D representation choice should be motivated At present there is no information on the motivation behind the selection of a 3D terrain representation as the novel representation type, other than its novelty. There seems to be broader context missing.

Why a 3D visualization? Why a map representation? What are the applications of understanding how to train people to use 3D terrain visualizations (what are the applications of 3D terrains in general?)? I have a person bias because this touches on my core research areas, but think the paper would benefit from some modest expansion here. Within visualization research, 3D representations (and 3D map representations specifically) are often skeptically viewed, and I think there needs to be some reasoning provided behind the choice to use them here. I think there are certainly military and emergency management applications that could be pointed out at the very least. People seek to use 3D terrain representations regardless of research controversy, so understanding how they can be better leveraged (e.g., trained) is a worthwhile endeavor. I provide citations near the end of my comments, but to summarize (my understanding of) 3D visualization controversy: 3D representations are somewhat controversial (or at least often-debated) within the information visualization and cartography communities because their usefulness historically has not matched their visual appeal. Generally speaking, 3D representations in/on 2D media (e.g., 2D non-stereoscopic displays) tend to be less effective than well-designed 2D representations unless the data is inherently highly "spatial" – e.g., a total brainscan or subterranean features. 3D representations on 2D screens suffer from distortion (e.g., perspective distortion) that makes positions, sizes, and shapes ambiguous (in real life humans have access to other depth cues available that eliminate most ambiguity), as well as occlusion. Provision often needs to be made for navigation that in a 3D space adds cognitive overhead that is not present in most 2D displays. I would consider this comment resolved if the authors could add some additional motivation within the introduction, ideally a paragraph summarizing related 3D terrain visualization research, but a more modest statement of motivation with an acknowledgment that 3D visualization is a deep/controversial would ultimately be adequate.

2. Examine research on "naïve" cartography for potential relevance to results The authors should review research on naïve visualization by Smallman and Hegarty et al. and consider it for inclusion in their discussion (and, if appropriate, introduction). The basic take-away from naïve visualization research is that users/participants tend to prefer "realistic" (e.g., 3D) displays that contain extraneous information for a task, yet also perform tasks less well with those displays compared to more simplified displays. Importantly, response time is sometimes particularly impaired. In a way, I see the results presented in this paper as a showing that practice with a task partly mitigates the performance detriment of "realistic" displays. In the study presented, the terrain stimulus was not task-relevant, i.e., for the purposes of the task it only served as visual noise. However, it appears that when the participants were trained on the task with an equivalent 2D representation, they were able to successfully able to "tune out" the terrain and focus on the task-relevant symbology, i.e., the dots or the ellipses. Interestingly, even with feedback built into the task (which is obviously effective for accuracy), participants without training seemingly persisted in being unable to ignore the terrain completely, hurting their response time. This strikes me as related to the findings of naïve visualization research, summarized below. The results seem to indicate that practice/pre-training appears to mitigate some of the loss of performance (with respect

to response time) that's present in tasks with extraneous (but perhaps aesthetically appealing) visual information. Specifically, I advise the authors to take a look at the following literature (with no obligation, implied or otherwise, that they cite it any in a revision). Full citations are included at the end of the review response document. Smallman, St. John, Oonk, and Cowen (2001): Response times were higher for identifying 3D symbols compared to 2D symbols, but similar accuracy – seems somewhat similar to the results here. Smallman and St. John (2005): This is a good digestible overview of naïve realism as a research topic and the issues of 3D representations on 2D displays. Hegarty, Smallman, Stull, and Canham (2009): Most relevant reference(i.e., if you look at one of the suggested papers, look at this one); a complex background map hindered accuracy of non-expert participants but increased the response time for expert and non-expert participants. Choice quote: "Both Study 3 in the present article and previous research (Canham and others 2007) show that adding terrain details to a weather map can mask more task relevant information and increase response time on a simple read-off and comparison task by 10% or more, while each additional irrelevant meteorological variable added between 5% and 10% to response time. Similarly, research that formally modelled the amount of visual clutter on different displays has shown that the speed of visual search increases systematically with additional variables, which add visual clutter (Rosenholtz, Li, and Nakano 2007)." (Hegarty et al., 2009, p. 183) Smallman and Cook (2011): Terrain-based task that is quite a bit different from the task in this paper, but has a good overview of research to that point and some good citations to cartographic/terrain visualization research that may be helpful. Also includes evaluation of spatial abilities as relevant individual differences. Hegarty, Smallman, and Stull (2012): This extends work form Hegarty et al. (2009) and has similar results; users performed worse with more "realistic" maps. I would consider this comment resolved by either an added discussion within the paper of this and/or related research, or a response by the authors explaining why this literature is not relevant.

3. Hypotheses should be more clearly and precisely communicated Overall the actual hypotheses are difficult to parse out in the current text and need to be better (i.e., more explicitly) defined and connected to other parts of the paper. I acknowledge that the analysis is exploratory but there needs to be better articulated logic underlying the experiment design within the introduction that is returned to within the results and discussion. Right now there is a lack of explicit mapping that would help tie the work together. Specifically:

- There should be a stronger, more explicit connection between the preceding theoretical discussion and the hypotheses being made.
- Generally, the hypotheses do not seem as cohesively linked to the experimental conditions/levels as I would like.
- There should be more direct references to the performance measures used, e.g., to response ratio accuracy.
- Response time was modeled, but did not seem explicitly referenced within the hypotheses.
- For the ease of reading, I would suggest making the hypotheses stand out, such as by

using some labeling scheme (e.g., Hypothesis 1, H1, etc.), so that they are more easily tracked through the rest of the paper.
- There is an inconsistency in terms used.
- Within hypotheses, the task complexity is referenced as "two estimates" and "four estimates" but later (e.g., in the figures), the terms used are "simple" and "complex" – I suggest making this more consistent or at least referenced, such as phrasing it as "two estimates (simple task) or four estimates (complex task)" – or something similar.
- Similarly, there seems to be inconsistent use of the term "practice" and "training" – the former being used within the hypotheses but the latter being used in the figures.
- There seems to be some fuzzy wording within what appear to be hypothetical statements, e.g., "should", "might" where I would expect an unambiguous if-then assertion.
- e.g., "the benefits of such practice might begin to break down under higher complexity." – this does not seem to be something practically measurable (e.g., how to define "break down"?).
- For example, within these statements, I would expect something more straightforward such as "If practice only reinforced heuristics and does not improve inference, then there will not be a (significant) difference in task performance between practiced and unpracticed participants in complex tasks.
- The hypotheses do not appear to be only indirectly referenced within the discussion, which makes it more difficult to evaluate that the claims set out in the introduction were validated (or not). In my reading, the authors had several somewhat overlapping hypotheses. I have written my interpretation of those hypothesis below. They were initially created within my own notes but provided below to the authors to hopefully provide some context for my above comments. I would appreciate concurrence/non-concurrence on whether they have been properly captured at least at a high level.

Practice

- H1: If practice improves speed and fluency of underlying inference (i.e., transferable inference), then practiced participants will perform better than unpracticed participants on a task using a different representation type but requiring the same underlying inference.
- H1A: If practice only improves visual processing (and not inference), then practiced participants will show no improvement over unpracticed participants on a task using a different representation type but requiring the same underlying inference.
- Question: Is this being made with respect only to the distribution visualization and not 2D to 3D? For 2D to 3D I don't think the tasks are distinct enough to support this claim. The representation and task are identical except for additional navigation and distortion (due to the terrain drape and 3D perspective).

Complexity

- H2: If practice in a lower-complexity (2 or 3 provided estimates) task improves transferable inference, then we expect higher performance for practiced participants in a high complexity (4 provided estimates) compared to unpracticed participants.
- H2A: If practice only reinforced heuristics and does not improve inference, then there will be NOT be a significant difference between practiced and unpracticed participants

in the degree to which their responses are in the extremes of one or more of the provided estimates.

- Note: Responses in the extremes of an estimate should indicate when a participant did not integrate knowledge of that estimate into their response. Transfer
- H3: If practice in simple tasks reinforced non-generalizable heuristics, then we would expect performance benefits for practiced participants in simple tasks in 3D, but not complex ones. I would consider this comment resolved if the authors make the suggested changes regarding the hypothesis presentation, or provide a response providing sound reasoning for not making those changes (e.g., I missed something, or better alternative was found). Additionally, a response regarding the validity of my interpretation of the hypotheses should be addressed, but it is not necessary to do so in detail.

4. The Figures need revisions for consistency and readability Overall, I would encourage modifications to the figures to increase readability, with some issues being larger than others. First, there are specific inconsistencies and error that need to be addressed:

- Trial time as 0-1 in some figures – should this be trial number? It's not clear if this is an artifact of the modeling, but is much more intuitive as trial number regardless (it's not strictly "time") e.g., Figure on page 11.
- Inconsistencies between the caption and the figure, e.g., "p(Zero)" within the figure on page 12 is not directly referenced.
- Regarding Figure 1, I would prefer having some more labeling embedded within the figure, e.g., referencing the type of representation and the number of independent estimates
- Regarding Page 11 Figure
- The line graphs are largely incomprehensible other than the "tails" for the simple trials, I would suggest these be reworked. Making them bigger would help, but probably ditching one of the credible intervals is necessary.
- "Starting" and "ending" – what are the definitions of these? e.g., first and last 10 trials?
- The other figures should be checked for these same issues. Second, I encourage the authors to make the figures more self-contained in terms of the information they portray. I understand this is something that varies in terms of standards and preferences, but I am of the mindset that figures should be able to largely stand on their own as communication devices without excessive reference to the caption or main text. Specifically, I advise the authors to:
- Add a title to make it clearer what is being communicated (especially with respect to other very similar figures)
- Add sub-labels that provide an explanation for labels such as task type/task complexity.
- I have assembled a quick example below:

I would consider this comment resolved if the authors correct the errors pointed out, and either make the suggested changes or provide reasoning for not making those changes.

5. Minor Issues

- The formatting of "i.e." is not consistent within the introduction – see paragraph 4 versus paragraph 7 (comma usage and italics)
- "Sand table" I'm familiar with as a military term, but may not be understood more widely, and a brief explanation of what is meant by the term in this context may be helpful (e.g., 3D tabletop map).
- Practice versus Training – this terminology should be consistent across the paper and within the figures. I recommend sticking to "Practiced" and "Unpracticed" groups since they were all given some training in order to complete the experiment.
- I found the terminology for the performance measures to be confusing and suggest the authors consider renaming them, and check they are referenced consistently.
- I recommend using common terms associated with performance in labels (e.g., score, accuracy, error) so that is clearer if higher or lower values are "better" – with a task it should be clear how performance differed and in what direction.
- For example, I suggest calling "response ratio" the "accuracy score"
- When referencing modeled parameters rather than direct measures, I think a simple parenthetical reference is sufficient, e.g., "performance metric score x (modeled by parameter y)" to save the reader having to generate an answer key of sorts between the metric and the model.
- Figures appear to be mis-numbered (oops!)
- There is a Figure 1 on page 4, and on page 11 (which should be 3, and it appears the numbering restarted there – text references seem to be to the original 1-7)
- Figure 1 caption has a formatting issue – looks like two different font sizes crept in.
- Adding a visualization of the raw performance data within the paper (rather than supplementary materials) may be helpful, for example a boxplot showing the raw performance figures (e.g., mean performance by condition), or a table of some summary statistics.
- I believe "2-d" and "3-d" should be "2D" and "3D", respectfully, unless the authors can provide a convincing argument otherwise – 2D and 3D are de facto standard terms for these types of representations.
- Section 2.2 references the task as a "game"- this seems out of place as there is no other reference to the task being a "game". Were participants cued to this being a "game"?
- In Section 2.3, it is stated that the stimuli distributions are "error" distributions – I believe "probability" distribution is more appropriate.
- Section 2.3 calls the stimulus a "realistic snowy terrain model" – I would simply call this a "terrain model" to avoid nitpickers like this reviewer having something to nit-pick. This is a very default looking Unity terrain to someone familiar with Unity, and "realism" is just too nebulous to be useful here.
- Section 2.5 – citations should be provided for the questionnaires if applicable, and they should be provided in the supplementary materials unless proprietary or otherwise restricted.
- Section 2.5.1, the second link is not formatted correctly; the url when clicked includes the closing bracket and so leads to an error page. It looks like this may be happening with the other OSF URL as well, but does not produce an error.
- Section 2.5.2 references a "Capsule Collider" – I believe this may be a Unity-specific term, but regardless it is an unnecessary technical detail (it is sufficient to state that

participant viewpoints were not allowed to pass within or through the virtual table).
- Page 8, Section 2.6.1, there appears to be some changes in the text spacing (line beginning with "Each of the parameters…")

Reviewer Name

Mark B Simpson

ORCID

0000-0001-9946-8161

Review #3

Completed: 18-11-2023 22:10

Recommendation: Accept Submission

Conflict Declaration

I declare that I have no known conflicts of interest with the authors.

Review

This paper addresses important questions about the impact of training on decision making with uncertainty visualizations and when/how that training can transfer to other contexts. There has been relatively little research in this area, so this paper makes useful contributions. Some of the findings were unexpected, which points to questions that could be addressed in future research. The paper is well-written and well-structured, so it is easy to follow. The claims are supported with detailed descriptions of the methodology and analysis as well as supplemental materials.

Openness

The authors have done a great job of supplying their materials. I personally am not familiar with the types of analyses they used in this paper, but the analyses are thoroughly documented. Given this transparency, it should be easy for readers to evaluate the analyses and the support for the authors' findings. I especially appreciate the authors' inclusion of links to their tasks.

Classification

Empirical Research - Quantitative

Recommendation

Minor Revisions

Revisions Requested

There were two things that I would like to see explored further in the paper. I am wondering if fatigue may have impacted the participants who completed the training prior to the transfer task. Please report how long the study lasted for participants in the training and no training conditions. The differing lengths of the experiments may help to explain why the untrained participants were performing slightly better than the trained participants at the end of the simple block.

Second, in prior studies that have compared ensemble and summary visualizations of uncertainty generally haven't had multiple overlapping distributions like those in this study. The overlapping scatterplots may have been more difficult to interpret than the ensemble visualizations used in prior work. I think it would be helpful to include some discussion of this in the paper. How much did the scatterplots obscure one another and how might that have impacted the participants' strategies and task performance?

Reviewer Name

Anonymous

ORCID

N/A

Metareview

Completed: 20-11-2023 17:15

Recommendation: Major Revision

Conflict Declaration

I declare that I have no known conflicts of interest with the authors.

Review

Thank you for your submission to the Journal of Visualization and Interaction. There is some consensus from the reviewers that your manuscript represents important contributions to the field. Each reviewer has provided a list of recommended revisions that will enhance the quality of the paper. We are categorizing this paper as requiring minor revisions, as all the requested changes can be addressed through writing modifications or with some analysis adjustments. We encourage the authors to address all the comments from the reviewers, and the main areas of improvement are summarized below.

Revisions Requested

1. **Research Background and Significance:** The authors should provide a more detailed explanation of the research background and its significance, ensuring a clearer

presentation of the study's context and relevance in the broader field. This includes incorporating research on naïve visualization, which suggests that users prefer realistic displays but may perform worse with them.

2. **Experimental Design Concerns:** Analyze the impact of viewpoint changes or address this in the limitations section. The authors should also implement additional analyses as suggested by the reviewers.
3. **Presentation Issues:** The paper requires improved clarity of writing, organization of content, and the visual presentation of data and figures.
4. **Exploration of Fatigue Factors:** The reviewer suggests investigating whether participant fatigue affected the results, particularly for those who completed training before the transfer task. Reporting on the duration of the study for different participant groups is recommended.
5. **Analysis of Overlapping Distributions:** The authors need to discuss how overlapping scatterplots in the study might have influenced participant interpretation and performance.
6. **3D Representation Justification:** Justification is needed for 3D terrain representation, including its motivation and applicability in fields like military and emergency management. The paper should also address the controversial nature of 3D visualization in information visualization and cartography.
7. **Clear Communication of Hypotheses:** The hypotheses require clearer and more precise communication, with a stronger connection to the experimental design and results. This includes maintaining consistency in terminology and using more explicit statements.

Addressing Minor Issues: Please fix minor issues, including formatting inconsistencies, unclear terminology, errors in figure numbering, and appropriate referencing of specific technologies or tools. Additionally, we apologize for the longer-than-usual review cycle. Unfortunately, a previously assigned reviewer discovered a conflict of interest very late in the process, necessitating the appointment of a new reviewer. Thank you for your understanding and patience in this matter.

Reviewer Name

Lace Padilla

ORCID

0000-0001-9251-5279