Journal of Eye Movement Research

Responses to the reviews of the manuscript

Tang et al.:
Eye-tracking study on solving science ordering problems

Dear JEMR Editors,
Thank you for the opportunity to revise and resubmit our manuscript, "Eye-tracking study on solving science ordering problems", which we have retitled "Eye movement patterns in solving science ordering problems". In line with your call for Major Revision, we have made extensive changes in the revised version.
This revision addresses all of the comments from the reviewers. We have also shortened the manuscript.
We believe that the changes made have strengthened the presentation of the research. We thank the reviewers for their work to improve this manuscript and for their constructive comments and suggestions.
An itemized listing of changes made in the manuscript follows. Responses to the reviews are in '[ ]' under specific comments from the reviewers.

REVIEW 1
c) Recommendation

Major Revision

2) Reader interest
a) Is the paper of current interest to a reasonable segment of the readership?

In general, science problem solving is an important field of basic and applied (educational) psychology and recent contributions made by eye tracking research have advanced this field considerably. The very specific nature of the investigated ordering problems limits the interest of the paper somewhat. The authors should emphasize where these problems are prominent and how their results can be generalized to other problem solving tasks.

[We have more fully explained why the ordering problems were used and emphasized where in science education these problems are prominent. This was done primarily in the Introduction.]

b) Is the paper likely to be used by other researchers and is the topic of the paper considered important?

There are a growing number of papers using eye tracking to study educational questions and I expect that future research could refer to the present paper. If the inappropriate interdependency analyses are dropped (see below) and the paper focuses on the problem solving itself, then it may make a valuable contribution. If other material is also removed/shortened (see below) the resulting length of the paper will be appropriate for its contribution.

[The interdependency analyses were drastically shortened (see below).]

3) Content
a) Is the paper technically sound?

Yes.

b) How would you describe the technical depth of the paper?

The eye tracking (ET) recording and analysis is kept simple. A standard AOI analysis using standard eye movement (EM) measures is conducted. The application of factor analysis to eye movement measures is unusual and inappropriate (see below).

I find that the authors could have used the ET data in a more advanced way to clarify, how participants process ordering problems (see recommendations under 5 below).

[Factor analyses were removed, as recommended.]

c) Does the paper make a contribution to the state-of-the-art in its field?

The strength of the paper is the analysis of phases and patterns in the problem solving process. The contribution could be stronger however, if the sequential ET data (phases, patterns) are analysed in greater detail (transitions), see detailed recommendations below.

[The ratios of fixational transitions have been added to the Results.]

d) Does the paper make adequate reference to earlier contributions?

Yes.

4) Presentation

a) Does the title adequately reflect the content of the manuscript?

No. The prefix "eye tracking study" makes no sense for JEMR. The authors should find a title that better emphasizes the content. I was not familiar with the term "ordering problems". Maybe a more general term exists? Title suggestion: Eye movement patterns and sequential organisation in solving [text-based] science [ordering] problems.

[The title was changed to "Eye movement patterns in solving science ordering problems".]

b) Is the abstract an appropriate and adequate digest of the work presented?

The abstract will have to be adapted to reflect the changes after a possible revision.

[The abstract was changed accordingly.]

c) Are the keywords well chosen?

Yes.

d) Does the introduction clearly state the background and motivation in terms of being understandable to the non-specialist?

The intro is somewhat underdeveloped/general. It needs to be more specific. In particular, an analysis of the cognitive demands that ordering problems pose to the subject is necessary. The authors have started with such an analysis (the three phases) but inappropriately hidden it in the materials section. Their analysis of the phases and the respective transition patterns is the important part of their work. It must be set up in the intro.

[Definitions of phases and previous work were provided in the Introduction.]

The authors should devote some text to clarify what their research is contributing to the state of the art and what the novel aspects of their research are.

[The contributions of this study to the state-of-the-art and the novel aspects of the research have been added or expanded in the Inroduction and Discussion sections.]

A substantial part is devoted to cognitive load; thus, the reader expects that the investigation/results speak to this topic which is not the case. There are only 3 problems with 6 vs. 7 choices. This is not a useful manipulation, and accordingly the authors fail to find load effects. Similar concerns apply to the manipulation of different learning media. These sections must be trimmed or dropped altogether.

[The discussion regarding cognitive load was reduced, as recommended, and informtion about different learning media in previous studies was added to the Introduction. In doing so, the discussion about multimedia was condensed in the Introducation.]

e) Is the paper well organised?

The paper follows the standard organisation of an empirical paper. Some material seems misplaced; see above: Discussion/definition of phases should be moved to introduction.

[Discussion/definition of phases was moved to the Introduction, as recommended.]

To pre-empt/summarize my main concerns and corresponding recommendations (see details under 5 below), we need less interdependency analysis of EM measures and more scanpath analysis.

[Interdependency analysis of EM measures was shortened, as recommended.]

The discussion of eye movement measures in the intro (p. 3) takes too much space and is confusing. In the data analysis subsection (p. 6) the authors pick up eye measures again and add more confusion. Instead of reviewing terminology, they should say clearly what measures they used and why; see also h) below.

[The discussion of eye movement measures in the Introduction was refined, and the repetive discussion in the data analysis subsection was removed. The measures used in the analyses were clearly stated (the reason for using these measures was provided in the Introduction). The measure " fixation duration" has been specified as "total fixation duration", see h).]

Measurements for the stimuli should be given in degree v.a. (not just in pixels).

[An accuracy of eye-tracking measurements in degrees was added.]

f) Relative to its technical content, is the length of the paper appropriate?

No, the paper is too long. With my editing/removal suggestions the paper can be cut down to half of its current size which would be more appropriate for the kind of contribution.

[The paper was shortened, as recommended.]

g) Is the English satisfactory?

Yes. The text is competently written.

h) How readable is the paper for somebody who is not a specialist in this particular field?

Very readable. However, there is a massive problem with central terminology which pervades the paper and makes it often incomprehensible. The authors' usage of the term "fixation duration" is wrong (and probably inconsistent). Fixation duration denotes the "duration of each individual fixation". On p. 6 The authors quote this definition correctly from the Tobii manual. Thus, fixation duration is typically of the order of 200 - 300 ms. However, throughout the results section and probably in the intro already the authors seem to mean something else. Their "fixation durations" are several seconds long (e.g. Table 4). It was never clear to me what the authors mean when they use this term. In Table 4 I suspect it means time spent in respective phases, thus it is not related to individual fixations. Also, it is not always clear whether "fixation count" refers to the (overall) number of fixations (which is always correlated with time on task) in a trial or to the number of fixations in specific AOIs (which could be completely uncorrelated with time on task). It is imperative that this is cleared up (see also my comments above under 4e).

[The measurement has been specified as "total fixation duration", which "measures the sum of the duration for all fixations within an AOI (or within all AOIs belonging to an AOI group)". The software used in this study, Tobii Studio 2.X, outputs mean fixation duration (in the order of 200 - 300 ms) and total fixation duration, and uses seconds as the unit of fixation duration.]

i) Is there unnecessary duplication of material in text, figures, tables?

The listing of R procedures in the appendix is unusual and unnecessary; remove.

[The listing of R procedures in the appendix was removed.]

Fig. 2 is expendable (because of Figs. 3 and 6).

[Fig. 2 was deleted, as recommended.]

5) Additional information
a) Specific comments and queries [please enumerate]

I have two main concerns.

(i) The results section (and the intro) focuses on the interdependence of several eye movement measures. I think this is inappropriate for several reasons. (a) Dependencies between these measures are well known (as the authors themselves acknowledge in the intro); some of them are trivial (e.g. total number of fixations and time on task) and not worth repeating. (b) Other dependencies however, vary depending on the task. For example, longer single fixation durations might indicate greater complexity of a stimulus or reduced discriminability, and are therefore not necessarily related to number of fixations. Thus, it is the responsibility of the researchers to select an EM measure that best suits the purpose (task) of the current study. They should select a small and meaningful number of EM measures and explain their decision to the reader. (c) My own preliminary analysis of the ordering problems is that they are text-based problems. This means that they consist of grammatically simple statement-like sentences that contain a large amount of specific terminology. The EM characteristics of such texts have been extensively studied in the reading literature. It is not appropriate to address them in the present context (problem solving). Another characteristic of the problems is that they involve mouse movements. This means that a substantial amount of EM will be devoted to the eye-hand coordination and does not reflect genuine cognitive processing which a priori renders several EM measures unusable, e.g. fixation durations in phase 2.

Recommendation: All the interdependence analyses (p. 7 -9, and respective sections in the intro) should be removed or radically shortened and moved to an appendix. The authors should work with time spent in specific AOIs (or alternatively number of fixations in these AOIs) and the order (pattern) in which the AOIs are visited.

[The interdependence analyses section was drastically shortened. The correlation matrix (Table 1) was moved to the Appendix. Total fixation duration was used as explained in h) above.]

(ii) The interesting and important part of this article is the analysis of fixational transitions (patterns) and the presence of different phases of the problem solving process. This part should be extended. This is the strength of this paper. I believe that a more thorough analysis of these data has the potential to reveal important characteristics of the problem solving process beyond the nice results that the authors have already obtained.

Recommendations:

Compute the (adjusted) ratio of the number of all fixational transitions (patterns) WITHIN the choice and steps areas relative to the number of transitions BETWEEN them for the three phases. This index should vary between phases; it would thus indicate that different processes are involved in these phases and that the division of the entire scanpath into phases is meaningful.

[The ratios in Phase 2 have been added to the Results section. Because there were very few "between" transitions in Phases 1 and 3, the ratios in these two phases were not computed.]

Analyze WHEN the first fixation of any of the steps areas occurs for correctly and incorrectly solved problems. This moment might indicate that participants have finished reading and start reasoning about the order. (Unfortunately, the steps are located to the left of the choices; therefore reading direction can interfere. Such flaws should be avoided in material construction.) Are there differences between successful and unsuccessful subjects?

[During Phase 1, almost all participants read the choices from the first to the last steps without fixational transitions between choices and steps. The exceptions were fixations on the question stem (Q) and the first step (1). Thus, this phase can be considered as a sole reading process without notable planning/reasoning about the order. This has been discussed in the Discussion.]

Patterns like "ABCD" denote reading/overview. Define other short "meaningful" patterns, e.g. fixations in Phase 1 that go from statements that follow immediately in the correct order. E.g. if the correct ordering of statements is "DFCEBGA", a "CE" or "EC" transition might denote planning rather than reading. The moment when those patterns emerge is important.

[Based on the results (Figure 5 in the revised manuscript), there were very few transitions during Phase 1. In other words, most participants read the choices naturally, i.e., from the first to the last steps. See Results and Discussion and the response to the previous recommendation.]

If incorrect solutions are not checked in Phase 3, is this because participants spent more time in Phase 2? Split the time spent per phase (Table 4) in correct/incorrect.

[The time participants were able to spend on a problem was not limited. As a result, it was unlikely that incorrect solutions were not checked in Phase 3 because participants spent more time in Phase 2. The total fixation durations per phase (Table 3 of the revised manuscript) in correct/incorrect have been split.]

Calculate a more refined solution score. E.g. assign points for each pair of correctly ordered

choices. Link EM patterns to those choice pairs. What are the EM differences that lead up to correctly vs. incorrectly ordered pairs.

[We added discussion about another grading system wherein partial credits were assigned, e.g., assigning points for each pair of correctly ordered choices. The results were similar (see p. 10 in the revised version.]

Specific remarks:

- • If the log regression coefficients in Tables 3 and 6 behave like normal correlation coefficients (ranging from -1 to +1) then the (significant) coefficients are negligible and practically meaningless. One more reason to avoid the dependency analysis.
- [The regression coefficients are scaled in terms of logs (ranging from -1 to +1). The exponential of the coefficient gives the odds ratios.]
-
- • The x axis label in Fig.5 must be “Problem”
- [The label has been changed, as recommended.]
-
- • p.2 the paragraph is too long
- [This paragraph has been revised and split into 3 paragraphs.]
-
- • p.4, participants subsection: “80%” of what?
- [It has been specified (eye-tracking ratio).]
-
- • P.4, materials subsection: How many static images were there?
- [The number (30) of the static images was added.]
-
- • p. 10.: What does Fig. 6 demonstrate exactly? What is it that a reader should note? As a more instructive alternative, consider an illustration of a scanpath from one participant which is coloured according to Phases 1 – 3.
- [The difference of gaze plots on the left (step) side between the correct and incorrect participants was added.]
-

• Fig. 7: The patterns should be ordered according to frequency; include pairs of transitions/patterns (see recommendations above).

[The patterns were ordered by length (of pattern substrings), and then by frequency, this is because we usually also are often interested in the length of pattern.]

REVIEW 2:

Review of "Eye-tracking Study on Solving Science Ordering Problems"

This paper discusses the findings of an eye-tracking study conducted to determine whether there is a difference in how students approach ordering problems. Participants were differentiated based on whether the problems were completed correctly or not. The methodology used is sound, although there are some minor issues, detailed below, which should be addressed.

One major concern I have is the absence of motivation for the study and what the significance of the results are. A motivation section should be added and the final results should be discussed in terms of the motivation and how they impact the reasons for the study. This will help to evaluate the impact of the paper within the scope of ordering problems. See my comments for more details.

[The motivation for the study was added to the Introduction. The significance of the results was discussed further in the Discussion section.]

Secondly, there are established methods which are used to evaluate the similarity of scanpaths which have not been used or referenced here. Please see my comment below for more information regarding this.

[We used established methods to analyze scanpaths patterns. The reference was added and the relationship between the method used in the referenced study and the one we developed was briefly introduced.]

Comments
Abstract - in the last sentence, it is stated the scanpaths of those who completed the problems correctly were more efficient than those who did not. How was efficiency measured, in other words, how is it determined what an efficient scanpath is? There is no mention in the scanpath section on an efficient scanpath. I suggest rewording the sentence in the abstract or clarifying this assertion in the scanpath analysis.

[A brief discussion of efficient scanpath/strategies was added (p. 10). The added part (ratio of fixaitonal transitions, p. 9) should also be related to this question.]

What is the motivation behind the study? In other words, why is it necessary to determine how students approach the problem? Why specifically these types of problems? Are there a large number of students who struggle with these types of questions, thus requiring some type of intervention or instruction based on the approach employed by the students who can correctly complete the ordering problem? A motivation section should be added to the article. The section entitled "Purpose of the Study" simply gives an overview of what the authors wanted to achieve and not why. The "why" question of the study must still be answered.

[We agree with the comments regarding the motivation for the study and have addressed this in the Introduction and Discussion sections, and have provided additional information about ordering problems.]

While this article is being submitted to an eye movement journal, it is customary to provide definitions for terms such as fixations and saccades as well as to distinguish between fixations and visits. A short paragraph explaining these should be added. The definition for visit and fixation duration etc came much later in the paper, after numerous references to them, please provide them earlier in the paper.

[The definitions of fixation and saccade have been added to the Introduction.]

Page 2 - Halmqvist should be Holmqvist

[The spelling was corrected.]

Page 3 - "...can click and move the mouse to rearrange the items on computer screen.". There should be an "a" between "on" and "computer screen"

[An "a" was added.]

Page 3 - A scanpath is simply constructed from the fixations and saccades using temporal sequencing.

[This definition has been added to the Introduction section.]

Page 3 - Research question 1 - research questions should not have a yes/no answer. Reword the question to avoid a yes/no answer.

[The question was changed, as recommended.]

Page 4 - remove the section where the authors give their opinion on what they envision the results would be. It could imply some bias on behalf of the researchers and should rather be removed so that the results can stand on their own to lead to the conclusions.

[The hypothesis section was removed, as recommended.]

Page 4 (Participants) - The wording seems to suggest that they were all enrolled in the same course, hence they were all on the same experience level? Please clarify this. In the discussion section, it is mentioned that participants may have had additional exposure to the content. Was this not controlled for in any way or evaluated using a pre-test questionnaire? This possibility should be discussed earlier when the participants are discussed and authors should indicate how they controlled for this or explain how it is not relevant within the scope of the study.

[This issue has been clarified in the Discussion section and mentioned in the Limitations section.]

Page 4 (Apparatus) - Why was a frequency of 60Hz used when the T120 is capable of 120Hz frequency?

[A frequency of 60Hz instead of 120Hz was selected because of large number of participants and long experimental time.]

Page 4/5 (Materials) - Was the content covered in the materials new to the students or was it taken from the content of a course they had already completed? Either way, the authors should clarify this and specify what the impact of this could be (see previous comment).

[This has been clarified in the Discussion section and mentioned in the Limitations section.]

Page 4/5 (Materials) - Was there a practice question to allow the participants to become familiarised with the way in which the ordering should be done (moving from right to left) or were they familiar with this type of interaction? Also, were all the participants provided with the options in the same order? Was the presentation order of the problems counter-balanced or are the questions of such a nature that counter-balancing is not required?

[The 17 problems were presented to the participants in the same order; this point has been clarified. There was no practice question because the drag-and-drop action was straight forward. We did not notice any problem with this type of interaction during the experiment.]

Was the difficulty approximately the same across questions since correct/incorrect was the only score given or should questions be weighted to control for difficulty (number of steps may not be sufficient as an indicator of difficulty)?

[The questions were independently considered to be of similar levels of difficulty by the content experts (two physiology professors); this has been added in the Materials section.]

What if the participant managed to place some of the steps in order and not all of them - should they not be given credit for this?

[In a second grading system, students were assigned partial credit if part of the step order in a problem was placed correctly. The results (e.g., regressions) were similar using the two different grading systems. This has been added to the Data Analysis, Results and Discussion sections.]

Page 4/5 (Materials) - Was it necessary to read all the steps before they could be placed in order or, for example, could one read the first choice and immediately know it was step 5? Please clarify this.

[According to the definition of reading and the results of Phase 1 depicted in Figure 5, most participants read from the first choice to the last. We have investigated the eye movement for each individual participant and found very few had cross gazes (between choices and steps) during reading.]

I asked this since a reading phase would be present. The next paragraph, however, addresses the phases which is very good. I would still, however, like to know whether all options must be read or if the answering phase can commence before all options are read? Maybe, perhaps once the (obvious) first step is encountered, the answering phase begins which might leave some "unfinished reading" which would then occur in phase 2?

[In this study, most participants finished reading the choices without looking at the steps (except for Step 1). See above. A concurrent *think aloud* could have answered this question more clearly.]

Page 6 - The acronym AOI is used the sentence before it is defined as Area of Interest. Normally, the very first instance is written out. Please correct (possibly since it is contained in a definition as something similar to "the duration of each individual fixation within an AOI" - when an AOI refers to an Area of Interest).

[This was revised as suggested, i.e., the very first instance of AOI has been written out.]

Page 6 - There are numerous articles available on scanpath theory (for example, Privitera & Stark, 2000; Noton & Stark). The section on scanpaths is relatively short in this paper, but cognisance should be taken of the fact that there is an entire field dedicated to the comparison of scanpaths. Analysis of scanpaths should be conducted using established methods or reasonable argumentation should be provided as to why it is not applicable to this study.

[The methods used are established ones, and we developed our own software tool based on it.]

Page 9 - "Problems 1 and 3 showed statistically significant", language must be corrected or additional phrase should be added to complete the sentence.

[This has been revised to "...fixation durations ... were significantly different".]

Page 10 - Authors refer to difficulty. What is the difficulty? Is it merely associated with the number of steps? On page 13, it is mentioned that student scores are indicative of difficulty. Is there not a more objective way of measuring difficulty?

[The questions were independently considered to be of similar levels of difficulty by the content experts (two physiology professors); this has been added in the Materials section.] We did not employ a more objective way of measuring difficulty.]

Page 10 - Reading would also cause multiple fixations within a single AOI (visits). Were scanpaths constructed using condensed scanpaths (successive fixations collapsed) or full scanpaths? I did not notice any repetitions in the provided scanpaths so perhaps they were collapsed? This is perfectly acceptable, it should just be mentioned by the authors.

[The word "collapsed. " was added, as recommended.]

Page 10 - The high level of similarity in Phase 1 and the lack of "jumps" is most likely due to the fact it is a reading phase.

[The reading characteristic in Phase 1 has been briefly discussed.]

Page 10/12 - The reason for the low incidence of patterns in phase 3 is unclear. It is stated that participants checked their answers step-by-step which is to be expected. However, if this is the case, scanpaths should be consecutive down the steps. Why is this not the case?

[We used a cutoff of percentage to identify the patterns and because not all participants (even in the correct group) completely checked their answers.]

Did the researchers do a pair-wise comparison. For example, if I completed question 1 correctly but not question 2 - how was this compensated for? Also, if there are participants where this happened, did their behaviour change from one problem to the next? This is perhaps beyond the scope of this paper but could be interesting as a follow-up.

[This within-subject comparison had been added to the Limitations and Future Work section.]

Page 13 - Regressions in reading are used as a measure to determine whether the reader is experiencing any difficulty. This study is not linear reading but I imagine that regressions could provide insight into the mindset of the participant during problem solving, either as a measure of checking or verifying information or possibly as an indication of planning or even that the participant is experiencing difficulty? Perhaps there are similar studies where the role of regressions are discussed within the cotenxt of ordering problems.

[The regressions in the manuscript is used to find difference in factors (e.g., media type and fixation duration in each problem/phase) between the correct and incorrect groups. The authors did not find similar studies where the role of regressions are discussed within the cotenxt of ordering problems.]

Page 14 - Referring to the sentence "Thus, it appeared that the participants set an inner clock to pre-assign time on each problem during the experiment even though they were instructed to solve the problems at their own pace. " Perhaps, 60% of total time is "standard" in terms of problem solving? Is there any literature which is available to determine if this is so? I doubt it has anything to do with an "inner clock" but is more likely general behaviour during problem solving, particularly it may also rely on the difficulty and scope of the problem which in this instance could have been similar. What is the general trend in terms of time distribution between phases 1, 2 and 3?

[The "inner clock" part has been removed.]

Page 14 - What is the significance of the study? What are the implications of the finding? These should link directly to the motivation for the study which must also be provided.

[The motivation for the study, and the significance and implications of the findings have been addressed in the Introduction and Discussion sections.]

Page 14 - Limitations and Conclusion sections should be swapped. The paper should end with the final conclusions of the study.

[The positions of Limitations and Conclusion sections were switched.]