Reviewer's report

Title: Characterization of medical students recall of factual knowledge using learning objects and repeated testing in a novel e-learning system

Version: 2
Date: 9 September 2014

Reviewer: Zineb Nouns

Reviewer's report:

Major Compulsory Revisions

Dear Authors,

thank you for submitting your study on the effect of studying and self-testing in an e-learning system on students self-reported knowledge recall. Even though I certainly do acknowledge the effort you undertake in this experimental study there are major constraints I would suggest you address before I would be able to judge the paper as suitable for publication. At this point I can not say whether the problems I see are due to the theoretical conception of your study itself or due to the structure of the paper that makes it hard for me to follow and I am not sure whether I correctly understand what you thought and did. The point I wish to address are in detail:

Background:

1. The introduction is not yet comprehensive. Here I would expect an introduction to the concept of spaced learning with references to the main literature in the field (i.g. Baddeley, Bjork & Bjork, ), test-enhanced learning (i.g. Roediger & Karpicke) in general. Followed by a description of the implications/impacts of this in the field of medical education including the use of technology enhanced practice. Right now it is difficult to rate the relevance of your study, as most of the Background part actually is not background but Methods : the entire description of the ALERT STUDENT tool itself does not belong here, this should be part of the „Methods“ section.

2. Quite some of the literature you cite comes from the background of Cognitive Load Theory yet the theory itself is not made explicit in the „Background“ nor elsewhere in the entire paper: please make clear what are the assumptions in this theoretical framework and how does that relate to your study and your decisions.

3. I am missing a clear research question and one or more hypothesis/hypotheses. Those should clearly relate to a (yet missing) sound theoretical background.

Methods :

1. It is not quite clear to me how you chose your participants: 98 where contacted and all volunteered? A rate of 100%? I need some more information: why did you decide to select 4th and 5th graders? How many are there in total? How was
data privacy handled? Who can see how the students perform in the self-assessments?

2. Regarding the self-reporting of recall accuracy: Did you use only self-reporting? What was your concept of recall-accuracy and basing on that, how do you argue the operationalization of this concept? How were the participants instructed to do this, what was their standard? Please specify on this.

3. Page 4, Line 84: I am quite confused at this point: Why is the pilot study reported here in the middle of the study design of the main study? As you have performed a pilot and used information from this, please describe this in the beginning of the methods parts making clear how the results from the pilot did effect your decisions on the design of the main study.

4. It is hard for me to follow the study-protocol itself by just reading your description. Maybe a visualization could help? I had to draw it in order to understand, I would not do this if it was a usual paper, I just did it as a reviewer.

5. Drawing the protocol of your design (if I understood it correctly) lead me to another conceptual question: the only difference between the experimental group and the so called control group was that the control group did not study? From this design I would infer, that your reasearch question was how the „studying“ intervention does alter the self-rating of recall? As your research question is not made explicit I can only assume this but if this is the case the text does not support this. And if this is not the case, the study-design does probably not fit. As you can see I am quite confused and I would therefore suggest to rephrase this part in order to make it understandable to the audience. And maybe draw a scheme of the design.

6. Page 5 Line 105: Please specify your blueprinting procedure! What makes it complete (comprehensive) and homogeneous?

7. Page 5 Line 109: Please specify: What was the rational behind the decision to remove the last three flashcards beyond the practical considerations (time constraint) you reported? What did the removal change in terms if validity, I mean was the content of the removed flashcards not necessary anymore or represented elsewhere?

8. Page 5 Line 117: Please make your decision for the method explicit: Why ANOVA? Did you consider multilevel models? From the data it looks like you performed some multilevel measurement but it is not clear if, what and why.

Results:

1. Page 6, Line 135: Please explain your local grading system as most of the readers might not be familiar with it. Please also include a standard in order to help the reader to understand whether the participants were good, average or poor performers and whether this is comparable to the entire students population.

2. By G-Score I assume you mean G-Coefficients? I find this part really hard to follow and –even more important – to judge and hence trust your results.

Discussion:

1. Page 7 Line 171: The sentence: « It was expected that the experiment group
would out perform the control group in terms of recall accuracy, at least on s1 » reads like it was A or THE hypothesis. Is this correct? If so, please outline it already earlier in the paper, i.e. at the end of the introduction. This would certainly enhance the structur of the paper.

2. Page 8 Line 192: you write: „This finding leads to the hypothesis that…“. I am not sure about the conceptional accuracy. To me it occurs as arguing the wrong direction as I had always hoped that hypotheses lead to findings and not the other way around. I mean of course findings might lead to further hypotheses to be adressed in future studies, but as you have not stated a hypothesis for the current study…..again, I am confused.

3. Page 8 Line 195-197: I regard the only self-report of recall accuracy a major constraint of this study. I would suggest to consider whether your study measured a well described effect (i.e. Simon, D. A. and R. A. Bjork (2001). "Metacognition in motor learning." J Exp Psychol Learn Mem Cogn 27(4): 907-912.): When learning in a blocked practice-setting the participants’ judgments (Metacognition/Knowing what you know) are unreliable and often overconfident. The current ease of a task makes learners feel that they learned more but in fact they didn’t when you actually test it objectively. I would argue in your study this effect might play a role as they rated their recall accuracy immediately after the study phase. And studying did only happen twice, which is not a classical „spaced-learning-design“ with many daily short studying session over weeks with a total of many study hours (i.e. Baddeley&Longman 1978: The influence of length and frequency of training sessions on the rate of learning to type. Ergonomics, 21, 627-635).

4. Page 8 Line 214: You state that there is a benefit of testing with study session in enhancing learning: Actually this effect is mainly shown for testing rather than testing + studying: In many of the studies on the testing effect, the testing seems to have more impact on learning than the studying: (See i.e.: Roediger&Karpicke: the critical importance of retrieval for learning. Science). This is why the phenomenon is called the testing effect and not the studying effect #.

5. Page 9 Line 231: This is an important statement. Students commitment and effort often matters much more than environmental/instructional design (even though this is quite hard to accept for us as curriculum/environment designers #). Please add literature to this statement (i.e. Artino, Pekrun). But again, the self-rated recall accuracy increase might be best explained by Bjork as mentioned earlier.

6. Page 9 Line 233: I do not understand the term « harder Materials ».

7. Page 9 Line 235: I am confused by this focus on self-assessment. I do not oppose the idea that self-assessment does add to fostering learning and knowledge retention, but it is by far not the only way to do so. I would like the authors to broaden their perspective here towards formative assessment such as i.e. progress tests or other feedback oriented assessments, remediation concepts, mentoring, portfolios, electronic feedback tools etc.

8. Page 9 Line 239-242: The causal relation between using ALERT and self-rating recall accuracy and learning in new settings assumed here is not
intuitive to me.....Do the authors refer to the quite complex idea of transfer? Transfer was not investigated in the study, was it? Maybe I get it completely wrong. You may consider clarifying this argument.

9. Page9/10 Line 247-249: You write: „It is not known whether study and self-assessment sessions are more effective than study sessions only. Other authors have shown that study activities not followed by assessment activities result in less knowledge retention than study plus assessment activities [34, 40].” To be honest, now I am lost: I thought that was your point? Now you state it's not known anymore. However, this is exactly the kind of discussion from the literature that I am missing in the „Background“ part. If you place it here it kind of counteracts your entire assumptions that lead to your study.

10. Page 10 Line 252: Yes, I already asked why you chose year 4 and 5 students? Maybe you had a good reason to do so beyond feasibility?

**Level of interest:** An article of importance in its field

**Quality of written English:** Acceptable

**Statistical review:** Yes, but I do not feel adequately qualified to assess the statistics.

**Declaration of competing interests:**

None