Author's response to reviews

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

Authors:

Tiago Taveira-Gomes (tiago.taveira@me.com)
Rui Prado-Costa (rcosta4540@gmail.com)
Milton Severo (milton@med.up.pt)
Maria Amélia Ferreira (mameliaferreira@med.up.pt)

Version: 3
Date: 30 October 2014

Author's response to reviews: see over
Dear Editors,

Dear Mr Christoph Nikendei,

Dear Mr Aldrin Ulep,

I am very thankful for the insightful revisions and suggestions provided by the referees. Most, if not all, revisions have been incorporated in the revised version of the manuscript, and have most undoubtedly increased its clarity and value.

In addition to the requested changes, some values were not correctly reported in the previous manuscript and have been now corrected, namely:

- Total number of questions in the notebook was 63, not 49
- Number of students in the sample was 96, not 98

Starting in the next page follows a detailed point by point response to the revisions and concerns raised by the referees. Reviewers are identified by same numbers previously used for identification. The italic sections are quotes from the original referee manuscripts, and the numbered points refer the actions that we considered relevant to meet the referees concerns. Most actions include additional comments for clarification or excerpts from the revised manuscript.

I hope that this version of the manuscript fully meets the concerns raised by the referees and thus may be judged as suitable for publication.

I am fully available to address any additional inquires that you or the referees find relevant.

Best Regards,

Tiago Taveira-Gomes
Reviewer 2

The study questions are described globally at the background section and could be further specified to demonstrate the underlying hypotheses. The questions should be concrete answerable by the evaluated data.

1. Clarify the study question and enumerate the underlying hypotheses making sure they can be answerable by the evaluated data

The background extended and a section named “Aims to this study” was created where the study question and the underlying hypothesis are made clear.

In this context Discussion, line 171-2 suggests that the author expected a group difference on s1, not essentially on s2. As the intervention group received a second learning intervention before testing compared to control it remains unclear to me why a potential difference would scale down.

2. Clarify why the learning intervention would scale down on s2

Discussion > Evolution of recall accuracy across sessions

…It was expected that the study-quiz group would out-perform the quiz group in terms of recall accuracy, at least on s1. Since the quiz task provides the learning materials as the correct answers to the OEQs and additional feedback at the end of the task, it has high learning value. Because we used a 4 point scale to grade recall accuracy, it was reasonable to consider the hypothesis that the quiz task provides enough learning value to master the content and thus expect both groups to report similar recall accuracy results…

The aspect how the described e-learning system could be curricular integrated needs to be further discussed. The text suggests that the e-learning tool is offered as an optional and add-on learning activity that is open to both, teachers and learners. This raises questions like who is responsible for the content and its maintenance, the medical school or each section, and would students spend time for an optional learning activity when their schedule is already busy as reported. This is an issue on which many previous e-learning activities have failed, after huge efforts for setting up.

3. Describe how the system could be curricular integrated

Discussion > Proposal for curricular integration

This new section covers the requested topic

4. Describe how students could integrate the system in their already busy schedule

Discussion > Potential implications to educators

…If educators take constructive action from such observations then a positive feedback cycle between student engagement and the success of the learning activity would be established. Because students know educators can take real-time action based on their progress, they engage more strongly in the learning activities. Stronger engagement will lead to better learning outcomes, that will lead to further tailored action by the teacher. Indeed, student engagement is the main driver of learning outcomes [80]. Providing tools that can foster such engagement is key to achieve successful learning [81,82]
Discussion > Potential implications to learners

The feedback that is thus formed between the quiz and the study task further promotes the spaced repetition of study and self-assessment sessions and can improve student engagement, the main driver of successful learning. This is even more important at a time where students need to define tangible goals that allow them cope with course demands [83].

For successful establishment of e-learning resources ‘blended learning’ approaches are widely described. Such use is implied in the text but could be discussed more explicit with regards to other published approaches which are similar. In this context the author should describe how ecological scenarios for the usage of the platform could look like to explain its surplus for the med school.

5. Discuss other published similar approaches regarding blended learning

Discussion > Proposal for curricular integration

In recent years multiple educational interventions have described the benefits of implementing blended learning methodologies in medical education, namely in radiology [85], physiology [19], anatomy [17] and others [86,87]. However, the design of these interventions varies widely in configuration, instructional method and presentation [79]. Cook asserted that little has been done regarding Friedman’s proposal [88] of comparing computer based approaches rather than comparing against traditional approaches [79]. The platform ALERT STUDENT intends to add value to the blended learning approach, through the collection of recall accuracy data, and prescription of a method that can be systematically applied in most areas of medical knowledge.

6. Discuss ecological scenarios of the platform usage usage for the medical school

Discussion > Potential implications to learners

Each flashcard holds the recall accuracy for each student for each assessment. Increasing spaced repetitions of study and quiz increase the available recall accuracy data. Since notebooks can be constructed using any available flashcard, it is possible to create notebooks that include flashcards for which recall accuracy is already available. Therefore, advanced notebooks requiring background knowledge can include an introductory section composed of the most relevant flashcards about the background topics. This implies that without previous contact with the advanced notebooks, an estimate of how well the student recalls the background topics is already available. This increases the value of learning materials by fostering reutilization and distribution of LOs between different courses, educators and students [56,57,59,84] and promoting educator and student engagement [81].

Discussion > Proposal for curricular integration

As an example, the platform could be used to deliver the learning materials and provide the study and quiz features, that would act in concert with MCQ progress tests during class. Educators could use information about recall accuracy and number of study and quiz repetitions to gain insight on the relationship between test results and student effort. That information would be relevant to help educators mentor students more effectively. Again, the information brought by recall accuracy could be helpful to tailor other instructional methods and thus drive student satisfaction and motivation…
When describing potential implications to the learner such as a learning control tool to optimize self-directed learning time, other existing methods herefore should be discussed with pro and cons, e.g. progress tests or just a simple MC-test, that could be implemented easily on an online platform.

7. Discuss ALERT STUDENT learning tools together with other existing methods such as progress tests, MC-test, progress tests or other feedback oriented assessments, remediation concepts, mentoring, portfolios, electronic feedback tools
   Please see point 6, Discussion > Proposal for curricular integration

Abstract, Methods line 12 and 13: consisted of

8. Correct the expression consisted of on Abstract, Methods line 12 and 13

Methods, study design line 75: The text suggests that the respond rate to volunteer was 100% what surprises (98 out of 98?).

9. Clarify the participant selection method and the response rate
   Methods > Intervention design
   …Ninety-six (n=96) students from the 4th and 5th grades of our school were randomly picked from the universe of enrolled students (approx. 500), and were contacted via email to participate one month prior to this study. Two students promptly declined to participate and two more students were randomly picked…

Methods, study design, sample size estimation: Which difference between groups did you expect to find with which power of testing? What difference between groups is relevant (not only statistically significant) when using subjective testing methods?

10. Clarify the power of testing
    Methods > Pilot study
    …The average percentage difference in recall accuracy between the two student groups was 41%. Finding a similar difference in mean recall accuracy before and after an intervention using the study and quiz tools was assumed to be a reasonable expectation. Thus, the sample size required to discriminate statistical significance under such circumstances was n=48, assuming a power of 80% and a significance level of 0.05…

11. Clarify what difference between groups is relevant when using subjective testing methods
    We would define a difference between groups to be relevant by the relative percentage of change for the recall accuracy scale. To make it clear, we changed the reporting of recall accuracy from the original scale (0-3) to a 0 – 100% scale. Through out the discussion in general, we discuss the differences in recall accuracy between groups in a way that clearly conveys that the difference found between groups (~40%) is relevant in the context of out findings. As also now discussed in the paper, we can only make a judgment about how relevant is a change in recall accuracy when we map it to knowledge:
    Methods > Adequacy of recall accuracy as a measurement of knowledge
...Assuming a relationship between both variables is found, it would also be relevant to understand how different degrees of recall accuracy map to different degrees of knowledge...

Reviewer 3

It was not easy to understand the research question and the concrete study. I suggest an illustration of the study design and examples of notebooks and flashcards and a screenshot of the study mode and the quiz mode — at least as supplement. If I understood it right participants had learned the learning objectives in the past (Golgi three years ago) during the normal curriculum (are there grades or achievements available?). They were randomized. Both group start in week 1 with a self-assessment session. This means that open ended questions are provided according to the Golgi notebook. I assume the participants had to compute their answer and were after that provided with the correct answer. As I understood only a part of the questions was displayed and at S1 and S2 different questions were displayed. Why was accuracy not assessed objectively by counting the correct answers? The students rated their accuracy by using a likert scale. Was a feedback given? So far I understood they did not have time and chance to restudy. Participants came back after one week and quizzed again in the standard group but studied in the experimental group and immediately after that quizzed as well and graded their achievement. S1 was repeated after another weeks (S2). I do not really understand the design. The design investigates whether self assessed achievement in a specific topic is better after studying the contents immediately before the assessment. What is the pedagogical theory behind? In my opinion this neither investigates retrieval of knowledge to foster learning nor spaced repetition. What is the concrete research question, which is the dependent and which is the independent variable, what is the intervention in a more abstract view (based on the scientific literature)? Therecall accuracy is not well defined in the paper. What does the number 0.79 o 1.87 tell us? Were the questions the same in S0, S1 and S2? I suggest to provide the contents and the questions as supplement.

12. Add an illustration of the study design
An illustration of the study was added as Table 1.

13. Add the contents of the notebook and flashcards used for the study as a supplement
The Golgi notebook was included as a supplement — the original one however was written in Portuguese.

14. Add screen shots of the study and quiz modes as supplements
A paper regarding the ALERT STUDENT platform was previously published, where all the features and rationale are explained. In addition that paper contains screenshots of the study and quiz modes. Since that material is part of the results of the previous paper, we made an explicit reference to that paper in the introduction, so that the screenshots can be easily retrieved from that paper.

15. Clarify whether there are grades available regarding the Golgi from the time students learned it

Methods
For the purpose of this work, content about the Golgi complex was designed using lectures from the Cellular and Molecular Biology class, taught in the second semester of the first grade.

Methods > Sample characterization

The Cellular Biology grade was assumed to be the grade that best estimated prior knowledge about the Golgi...

16. Clarify how did the different questions vary and how they map to the presented content on both sessions

Methods > ALERT STUDENT platform

This new section clearly explains the different questions

17. Clarify which are the dependent and independent variables of the study

Please see point 1

18. Clarify the definition and the measurement of the recall accuracy

Introduction > Self-assessment and the ALERT STUDENT Platform

... The platform implements test-enhanced learning in the form of quizzes. These are composed of sets of OEQs about each of the LOs. The questions are meant to stimulate students to recall learned information, and therefore enable the measurement of JOLs. Typically, JOLs can be estimated as the prediction of the learner about how well it would recall an item after being presented the item [61]. Numerous methods exist to assess JOLs for different purposes [62]. The cue-only JOL, a method where the student must determine the recall of an item (in our case a LO) when only the cue (the OEQ) is presented at the time of judgment [62], is of particular interest to us. We extend this type of JOL to define a measurement named ‘recall accuracy’. The recall accuracy is similar to the cue-only JOL because after being presented the cue and trying to retrieve the target, the student is presented the LO that contains the target. The process of measuring recall accuracy corresponds to the immediate feedback stage employed on test-enhanced learning approaches. This approach maximizes the potential of LOs and the OEQ to serve as learning material, recall cue and recall feedback...

19. Clarify how participants were instructed to check the correctness of their answer and how that relates to recall accuracy ratings

In addition to point 19, please see

Methods > Intervention design

...Before each task students were instructed on the purpose of each task and the researcher exemplified each of the tasks in the system. Students performed each task alone. Doubts raised by the students concerning platform usage were cleared by the researcher...

20. Clarify what kind of feedback was given during self assessment and study sessions

The only feedback provided during the sessions was the feedback from the system for the quiz task. That feedback is described in the new section

Methods > ALERT STUDENT platform
Test-enhanced learning is achieved through the quiz mode, a complementary environment where retention of flashcard information can be self-assessed through recall accuracy using the OEQs as cues. Active recall is graded for each question using a 4 point Likert scale (0 - no recall, 1 - scarce recall, 2 - good recall, 3 - full recall). On every quiz session, the system picks one OEQ for every piece of information on every flashcard. OEQs are displayed one at a time. In case there is more than one OEQ for an information piece, the system picks one OEQ that has not yet been graded. When all the OEQs have been graded for a given information piece, the system picks the OEQ with the lowest recall accuracy. At the end of a quiz mode session, the student is presented the set of flashcards and OEQs for which recall accuracy was 0...

21. Clarify the rationale to perform the self assessment after the study sessions

This topic is addressed in general in the introduction and further clarified in the beginning of the discussion.

The experimental group has a longer time on task and additionally is provided with the answers of the assessment questions immediately before the assessment and therefore performs better than the group that was never confronted with the learning objectives in a controlled manner. Please explain more specifically what novel insights this result reveals and which evidence based conclusions we can draw for daily instruction from this result?

22. Describe the novel insights and conclusions given from the differences in recall accuracy between the assessment vs study + assessment

This topic is discussed in the sections:

Discussion > Evolution of recall accuracy across sessions
Discussion > Adequacy of recall accuracy as a measurement of knowledge

And is summarized in the study conclusion:

...We found that the quiz task alone led to a modest increase on recall accuracy, and that the study-quiz task had high impact in recall accuracy. The session effect was the main determinant of recall accuracy on the study-quiz group, and the flashcard and participant effects determined most of the increase in recall accuracy in the quiz group. We concluded that recall accuracy seems to be linked with knowledge retention and proposed further investigation to ascertain the nature of this relationship. Recall accuracy is an easily collectible measurement that increases the educational value of LOs and OEQs. In addition, we have discussed the educational implications of providing real-time recall accuracy information to students and educators, and proposed scenarios in which such information could be useful to deliver tailored learning experiences, assess the effectiveness of instruction, and facilitate research comparing blended learning interventions...

The authors again and again refer to the published research of test-enhanced learning. The authors should more precisely describe the influence of this theoretical background on the study design. Do they focus on knowledge acquisition or knowledge retention? As far as I understand the study investigates learning by repeated testing with feedback versus learning by repeated studying followed by repeated testing with feedback. The most published studies were focused on retrieval of knowledge versus acquisition of knowledge for knowledge retention. This does not compare to the presented study. Therefore I suggest to provide a more specific description of the theoretical background and a more precise description of the
lack of knowledge. Based on this I would be happy about a operationalized research question (see 1) and an illustration of the study design with concrete examples (see 1).

23. Describe the influence of test enhanced learning on the study design with precision
   This topic is now debated in detail in the section
   Introduction > Test-enhanced learning

24. Describe the theoretical background and the lack of knowledge regarding the research question with detail
   This topic is also generally answered by the revised introduction

25. Clarify the study question and enumerate the underlying hypotheses making sure they can be answerable by the evaluated data
   Please see point 1

26. Add an illustration of the study design
   Please see point 12

Figure 4 needs to be improved. The legend should clearly explain what the graph shows. Without the above mentioned information it is not possible to adequately comment on the results and the discussion section at this time point. With the imbalance of time on task that is systematically induced by the design it seems almost impossible to interpret the results. This is really a major concern. The other named limitations are less interfering.

27. Improve Figure 4 and clarify the legend
   The legend and the figure have been improved

**Reviewer 4**

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.e. Baddeley, Bjork & Bjork, ), test-enhanced learning (i.e. 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.

28. Describe the spaced learning concept in detail, its impact in medical education, and reference the suggested bibliography
   A section in the introduction has been created to present this topic in detail and its impact in medical education.
   Introduction > Spaced repetition

29. Move ALERT STUDENT tool description to the methods section
   The description of the platform features that were pertinent to the study design were moved to the methods section, under
   Methods > ALERT STUDENT platform
However, because ALERT STUDENT is a platform that was mentioned in another paper, and is important to introduce the theoretical background in regards to recall accuracy, we kept a mention in the introduction subtopic:

**Introduction > Self-assessment and the ALERT STUDENT Platform**

In addition that reference explicitly states that previous work regarding the platform, that can be useful to understand the system more fully (though not necessary to understand the current work)

*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.*

**30. Clarify the relationship between cognitive load theory with the study design**

Cognitive load theory is most important to justify other features of the platform that are not relevant to the current study. Therefore, a reference was made to this theory in general for medical education, and the topics of spaced repetition and test-enhanced learning, that are pertinent to this work, were explored more fully.

Addtionally please see, points 23 and 22.

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

**31. Clarify the study question and enumerate the underlying hypotheses making sure they can be answerable by the evaluated data**

Please see point 1.

*It is not quite clear to me how you chose your participants: 98 where contacted and all voluntered? 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?*

**32. Clarify the participant selection method and the response rate.**

Plase see point 9.

**33. Clarify the rationale for selecting 4th and 5th year medical students**

**Methods > Pilot study**

*The 4th and 5th year students knowledge was assumed to correspond to a low recall accuracy about the Golgi, and was expected to represent the mean recall accuracy of a similar student sample before the research intervention. 2nd grade medical students knowledge was assumed to correspond to a high recall accuracy about the Golgi, and was expected to represent the mean the recall accuracy of a student sample after the research intervention.*

**Discussion**

*It was unclear what difference to expect in terms of recall accuracy between groups and between sessions. We selected a basic science topic and 4th and 5th grade medical students, in order to maximize the odds of a low degree of prior knowledge.*
We chose the Golgi Complex because the majority of the curriculum does not build directly on this concept, and thus it was likely a forgotten topic. This was important because the lowest the a prior knowledge before our intervention, the smaller student sample would be required to discriminate significant differences in recall accuracy during the study sessions, thus rendering this study feasible.

34. Clarify how was data privacy handled and who had access to the student self assessment information.

Methods (last paragraph)

…Collected data was analyzed in an anonymous fashion. It was not possible for the researchers to identify the students during any phase of the data analysis…

Each student had indirect access to their recall accuracy that because that was used to compute the flashcard priority cue:

Methods > ALERT STUDENT platform

Spaced-repetition tools are made available through a study mode feature …enriched with … a flashcard study priority cue based on personal recall accuracy from corresponding OEQs (open ended questions).

35. Clarify the definition and the measurement of the recall accuracy.

Please see point 18

36. Clarify whether only self reporting was used to measure recall accuracy.

Recall accuracy is a judgement of learning that is collected through self report only. In addition please see point 18.

37. Clarify how were participants instructed to check the correctness of their answer and how that relates to recall accuracy ratings.

Please see point 19.

Page 4, Line 84: I am a 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.

38. Move the pilot study section to the beginning of the methods section and clarify its effect on the study design

The pilot study was refactored into a clear subsection and was moved to the beginning of the methods, just after describing the platform.

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.

39. Add an illustration of the study design

Please see point 12

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 the control group did not study? From this design I would infer, that your research 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.

40. Clarify the study question and enumerate the underlying hypotheses making sure they can be answerable by the evaluated data

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

41. Clarify how did the different questions vary and how they map to the presented content on both sessions

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?

42. Clarify the rationale behind the removal of the last three flashcards

The flashcards were removed purely because of the time constraint. The time constraint had to do with the laboratory schedules that made in unfeasible to design an intervention taking more than 20 mins for the study task.

The remaining flashcards did not contain any information regarding what was written in the three flashcards that were removed, nor were they changed it. Because the flashcards were presented in an incremental order, the removed flashcards were not necessary to better understand the remainder of the content.

In addition, please see point 38

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.

43. Clarify the rationale to use ANOVA

Methods > Statistical analysis

...In order to characterize the changes in recall accuracy across sessions, we used univariate repeated-measures analysis of variance (ANOVA). Groups were used as between-subjects factor. Session and flashcard were used as within subject factor.

44. Clarify the rationale for the use of multilevel measurements

Methods > Statistical analysis

In order to estimate the variance components for the recall accuracy for both groups, a random effects model was used and the flashcard, the session and the student were
used as random variables. The estimation was performed using the Restricted Maximum Likelihood method.

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.

45. Describe the school local grading system and map it to a standard easy to interpret

The scores from the exams were converted to a 0-100% scale so that they are easy to understand.

- Average course grade was 68%, and the average Cellular Biology grade was 64% - equivalent results for the student population were 65% and 62% respectively, representing a fair score.

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.

46. Clarify the rationale for the calculation of the G-Coefficients and explain its importance

This topic is now further clarified in:

Methods > Statistical Analysis
Discussion > Recall accuracy components of variance

Page 7 Line 171: The sentence: « It was expected that the experiment group would outperform 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 structure of the paper.

47. Clarify the study question and enumerate the underlying hypotheses making sure they can be answerable by the evaluated data

Please see point 1

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 addressed in future studies, but as you have not stated a hypothesis for the current study…again, I am confused.

48. Rephrase statement at Page 8 Line 192 in order to be less misleading

49. Clarify the study question and enumerate the underlying hypotheses making sure they can be answerable by the evaluated data

Please see point 1

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).

50. Discuss the extent to which the self report is a limitation of the study based on the suggested bibliography
This point has resulted mainly in the enrichment of the sections:
Discussion > Evolution of recall accuracy across sessions
Discussion > Adequacy of recall accuracy as a measurement of knowledge
And to a lesser extent:
Discussion > Recall accuracy components of variance
Introduction > Self-assessment and the ALERT STUDENT Platform

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.

51. Discuss the testing and study impact on learning based on the suggested bibliography
Please see point 50

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.

52. Improve the statement a Page 9 Line 231 with the suggested literature
Discussion > Potential implications to educators
...Because students know educators can take real-time action based on their progress, they engage more strongly in the learning activities. Stronger engagement will lead to better learning outcomes, that will lead to further tailored action by the teacher. Indeed, student engagement is the main driver of learning outcomes [80]. Providing tools that can foster such engagement is key to achieve successful learning [81,82].

Page 9 Line 233: I do not understand the term "harder Materials".

53. Rephrase the term at Page 9 Line 233 to prevent misleading interpretation
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.

54. Contextualize ALERT STUDENT learning tools with other existing methods such as progress tests, MC-test, progress tests or other feedback oriented assessments, remediation concepts, mentoring, portfolios, electronic feedback tools

Please see point 6

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.

55. Clarify the statement on Page 9 Line 239-242 to prevent misleading interpretation

Discussion > limitations and further work

The system works around factual knowledge, therefore it is only useful in settings that require acquisition of such knowledge. Complex competences such as multi level reasoning and transfer cannot be translated in terms of recall accuracy…

Page 9/10 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 its 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.

56. Clarify statement on Page 9/10 Line 247-249 to prevent misleading interpretation

This statement has been completely rewritten and its idea transposed to other statements on:

Discussion > limitations and further work

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?

57. Clarify the rationale for selecting 4th and 5th year medical students

Please see point 33

Reviewer 5

The Introduction/Background is too cursory. Indeed, key findings from the literature are mentioned but they are reported without theoretical background, framework, or context. The only justification for considering “spaced-repetition” and “repeated-testing” is that they “emerge” from the literature. The aim of the study (“to demonstrate”) is in a certain way misaligned with the design of the experiment (“self-assessment” vs. “self-assessment + study”).
58. Describe the theoretical background and the lack of knowledge regarding the research question with detail
   Please see point 1

59. Rephrase the aim of the study to be aligned with the design of the experiment
   Please see point 1

The entire section on “ALERT STUDENT” should be – in my opinion – moved to the methods section because this platform is a tool, an instrument in their study.

60. Move ALERT STUDENT tool description to the methods section
   Please see point 29

I found it very hard to follow the METHODS section. It jumps between several pilot studies, study design, content design, and (descriptive) statistics. The dependent variables are sometimes ill-defined and terms that seem to have a special use (e.g. “flashcards”) are introduced without clarification. I can guess how the flashcards looked like but a supplement/figure would be really helpful here. Please consider moving the reports on pilot studies to a supplement. Was the study conducted in a laboratory setting? Is this meant by “in-person session”?

61. Clarify which are the dependent and independent variables of the study
   Please see point 1

62. Introduce in a clear way any terms that denote a special use, such as the term flashcards
   The special terms have been consistently named, introduced and converted to italic style.

63. Move the pilot study section to a supplement
   Because the pilot study in our opinion is important to explain the study method, we decided to take advice from referee 4 requesting to move it to the top of the methods section. Please see point 38.

64. Clarify the effect of the results of the pilot study in the main study design
   Please see point 38

65. Add an illustration of the study design
   Please see point 12

66. Clarify that the study was conducted in a laboratory setting by rephrasing the "in-person" term

The “EVALUATION”-section reports details of the planned statistical analyses. I would suggest re-naming it into “Statistical Analysis” or something similar – but this is a minor issue. The main issue here is that it lacks detail. I don’t understand why one would be interested in including “flashcard” as a fixed effect into the analyses. Is there a rationale for this? Otherwise it might be more appropriate to account for this source of variation by including it as a random effect in a mixed-effects/"multilevel" model (indeed it is included as a random effect in the G-
I doubt that it would change the results drastically but the rationale for the choice is not explicit. The rationale/background of the calculation of G-Coefficients is not explicated (absolute/relative error?). Who should agree on what? And why is this important? Is it really intended to “determine the reliability of the variance”?

67. Rename the evaluation section to statistical analysis

68. Clarify the type of effect used in the analysis for the flashcard variable

   This is made clear in the section

   Methods > Statistical Analysis

69. Clarify the rationale for the calculation of the G-Coefficients and explain its importance

   Please see point 68

   Please see point 46

When estimating variance components it is common to mention which software and procedures were used (ordinary least squares, maximum likelihood, Bayesian procedures,...) as results may differ as a function of the estimation algorithms (cf. eg. Crossley et al. Med Educ. 2007 Oct;41(10):926-34.). I would also find it informative to know which software was used for the other calculations.

70. Describe the software and procedures used for the estimation of variance components and other calculations

   Please see point 68

From the METHODS section I had a mixed impression on the statistical procedure and this is also true for the RESULTS section: On the one hand the authors have conducted a set of rather elaborated analyses. On the other hand only a fraction of the results is reported which is a severe shortcoming. Indeed, there are – despite p-values – literally no results of the ANOVA reported (Df’s, Mean Squares/Sum of Squares, F-values, Eta-Squared, …). Thus, in it’s current form, the manuscript does not adhere to the relevant standards for reporting. Such details are needed to evaluate the soundness of both analysis and data.

71. Clarify the ANOVA result values that were obtained

   Results > Recall accuracy characterization

   … Regarding the ANOVA, the session and group Df’s equaled 1, Sum square/Mean square difference values were 56.5 for the session, and 23.5 for the group. F-values were 292.2 for the session and 121.2 for the group. Eta-squared values were 0.32 for the session and 0.27 for the group…

There are other details that are not reported on a rational. Why are average course grades important in this context? Why are study resources relevant? Why is Figure 1 presented? Why is study duration reported for the experimental group?

72. Clarify the rational to include the students average course grades and to include the students most used study resource

   Methods > Sample characterization
In session s0 both groups filled a survey to characterize the student sample. Measured factors were gender, course year, preferred study resource for Cellular Biology, computer usage habits, Cellular Biology grade, mean course grade, and average study session duration during the semester and during the exam season. The Cellular Biology grade was assumed to be the grade that best estimated prior knowledge about the Golgi. These factors were added to characterize the study sample and assess eventual dissimilarities in the sampling of the two groups.

73. Clarify the rational for including figure 1

Methods > Statistical Analysis

…In order to estimate the agreement on the flashcard component its specific G-coefficient was calculated. A D-Study was performed to characterize the agreement on the flashcard component for different student and session counts…

74. Remove the study duration reported as it is not directly related to the research question

The DISCUSSION gives a summary of the findings and their relation to the introduction and literature. However, it also reports additional results and even introduces new expectations. Please consider moving such information to the results and the introduction, respectively.

75. Move any additional results referenced on the discussion to the results

76. Move any new expectations introduced in the discussion to the introduction

Although being an interesting issue - the conclusions regarding affective/motivational influences on learning are, in my opinion, not supported by the data. As far as I understood the manuscript there is no information on such factors available in the current study.

77. Clarify the conclusions regarding affective / motivational influences

Please see point 50