Author's response to reviews

**Title:** Benefits of a physician-facing tablet presentation of patient symptom data: comparing paper and electronic formats

**Authors:**

Daniel Glaser (daniel.s.glaser@gmail.com)
Sanjula Jain (sanjula.jain@rice.edu)
Philip Kortum (pkortum@rice.edu)

**Version:** 3  **Date:** 17 April 2013

**Author's response to reviews:** see over
April 17, 2013

To the Editor:

We are pleased to resubmit our revised paper entitled, “On the benefits of a physician-facing tablet presentation of patient symptom data: comparing paper and electronic formats” to *BMC Medical Informatics and Decision Making*.

We have tried to carefully and thoughtfully address each of the reviewer’s suggestions in this revision, and believe that the paper is significantly better for these efforts. Our responses to each of the reviewer’s concerns are included below.

Sincerely,

Philip Kortum, PhD                 Daniel Glaser, Ph.D.                 Sanjula Jain

Rice University
Department of Psychology, MS-25
6100 Main Street
Houston, TX 77005
713-348-4813 (voice)
713-348-5221 (fax)
pkortum@rice.edu
RESPONSE TO REVIEWERS

REVIEWER 1

1.1) Electronic devices are a major subject of the study and screen size is the reason for experiment 2. Please specify the used devices. What iPad, what iPod, who is the manufacturer, what screen size and what resolution do they have? And on the other hand for the paper form: what size has the paper, how many pages on how many sheets does one MDASI survey comprise?

RESPONSE: We have described the details of the devices and the paper in the materials section. In short, the iPad was an Apple model MC349LL, which has a 9.7 inch screen with a resolution of 1024x768, running iOS 4.3. The iPod was an Apple model MC086LL, which has a 3.5 inch display with a resolution of 320x480, running iOS 4.0. Each MDASI survey fit on one 8.5x11 sheet of paper which has also been clarified in the text.

1.2) I am not sure how many MDASI surveys have been used in this study. For each mode of presentation (e.g. iPad vs Paper) it seems that each participant has to deal with a test set of 80 surveys (20 fictitious patients, each of which having 4 surveys). In the Section „Method: Experiment 1 – Paper vs. iPad“ „Materials“ paragraph 1 is stated: „…two twenty-patient sets were prepared for both the paper set and iPad application.“ Further down in the Section „Method: Experiment 1 – Paper vs. iPad“ „Procedure“ it is stated that „After all items were answered, participants […] then completed the same six questions on a different data set.“ To put it in a question: How many „different“ data sets have been used in this study? I suppose two (that makes 160 surveys). If i am right please phrase it more distinctly, for example like this: „…two different twenty-patient sets were prepared for both the paper set and iPad application.“ and „After all items were answered, participants […] then completed the same six questions on the other data set.“ If I am totally wrong, please clarify.

RESPONSE: There were indeed 2 fictitious patient data sets. A participant would use one dataset with one presentation format, and then another dataset with the other presentation format. We used two different data sets so that participants couldn’t just use their answers from their just-completed task. We have clarified this in the text by saying “Since participants were instructed to answer patient-related data using both the MDASI paper survey and the iPad, two different twenty-patient data sets were prepared. A participant would use one data set with one presentation format, and then another comparable data set with the other presentation format.”
1.3) Experiment 1 Results
The reported t statistics [t(17)] and F statistics [F(1,16)] imply that you have used the data of 18 participants and not of 17. Please explain this.

RESPONSE: There were indeed 18 participants, and we have correct this error.

1.4) Results general
The response accuracy and the SUS are on an ordinal scale (For accuracy there are different types of questions with different difficulties summed up and divided by the number of questions or for SUS a sum of 10 Likert scale questions multiplied by 2.5. In both measurements participants can only achieve discrete values). The correct way to present such figures is to use the median as the central tendency together with either the range (i.e. minimum - maximum) or the interquartile range (IQR = 3rd quartile minus 1st quartile) as a measure of the spread and not the mean with the standard deviation.

RESPONSE: As other reviewers have suggested, we have eliminated the figures from the paper. Median and IRQ are now reported in the results as well.

1.5) As a consequence of 1.4) when comparing two measurements that were described by the median nonparametric statistics should be used to assess the difference. In this case with the Wilcoxon signed ranks test instead of the paired t-test.

RESPONSE: We redid the analyses using the Wilcoxon test for each of the studies. The overall results did not change as a consequence of performing this different analysis.

1.6) Results, Task completion time
If the mean is suitable to report the time please display it with the standard deviation and not with the standard error (since you are describing the variation of different participants, each measured once and not the precision of one participant measured again and again). The mean is suitable to describe the time if the values do not deviate much from a normal distribution. You can assess this by the absence of extreme values or if the mean and the median (nearly) coincide. The paired t-test is appropriate to test for differences if the difference does not deviate much from a normal distribution. If there are extreme values of time measurements and the mean and the median differ much, I suggest to use the median with the range (or interquartile range) to describe the task completion time and again to use the Wilcoxon signed ranks test.
RESPONSE: We have switched to the Wilcoxon test, and now report SD rather than SEM.

1.7) Experiment 1 Results
An ANOVA is inappropriate to assess the relationship between the self-rated expertise and accuracy. Depending on how expertise was measured (see 2.3) I suggest the following methods (in order of explanatory value): In the best case you have many levels of expertise (e.g. scorepoints) because then it is suitable to calculate a Spearman correlation coefficient to find out whether there is a monotonic trend between expertise and accuracy (either the more expertise the more accuracy or the more expertise the less accuracy) or not. If you have „some“ levels of expertise (3 or a few more) I suggest to either use the Jonckheere-Terpstra test to show whether there is a trend or not or still to use the Spearman correlation coefficient. If expertise has only 2 levels (high, low) you can test whether the high expertise group has significantly different accuracy values than the low expertise group by using the Mann-Whitney-U test (sometimes referred to as Wilcoxon rank sum test).

RESPONSE: Because of the relative homogeneity of the expertise data (due to the population selection), and a lack of significant results, we felt it might be best to simply remove this from the paper for clarity and readability.

2.1) Section „Method: Experiment 1 – Paper vs. iPad“, „Participants“
Please further describe and explain the „17 undergraduates“. What branch of study? In the Introduction (Paragraph 2) when describing experiments 1 and 2 you denote them as physicians and in the Section „Physician Interviews“ (last paragraph) you are phrasing „for the purposes of assisting physicians“. On the one hand they are opposed to „medical professionals“ on the other hand 8 of the „medical professionals“ themselves were physicians.

RESPONSE: The undergraduate participants were drawn from the general student population. We have added this clarification. The introduction paragraph was not meant to imply that we used physicians or medical professionals for experiments 1 or 2. However, upon rereading it is clear that his could be an easy interpretation. For clarity, we removed the word „physician“ in the introduction when used in the context of experiment 1 and 2 and replace it with a more generic description of „a user who might interpret this data“. 
2.2) Method: Experiment 1

“Experiment 1 compared the common usability metrics of efficiency, effectiveness and satisfaction … “. You can’t compare „satisfaction“ because the SUS was not applied on the paper form. Either describe more detailed what you compared or state that „Experiment 1 assessed the common usability metrics of efficiency, effectiveness and satisfaction …“

RESPONSE: Indeed, we did not collect SUS measures for the paper version. The sentence has been rewritten to say “Experiment 1 compared the common usability metrics of efficiency and effectiveness [13] of the paper form of the MDASI and the electronic interface as implemented on the iPad across a number of representative tasks.”

2.3) Section „Method: Experiment 1 – Paper vs. iPad“, „Materials“ paragraph 1

Please describe the „general background & technology survey“. How is the „participants’ expertise level“ scaled? Is it a yes/no, scale points, low-middle-high assessment? The result of the ANOVA in experiment 1 implies that you have measured it in 2 levels or you have reduced the measurements to 2 levels.

RESPONSE: As noted in the response to Reviewer 1 comment 1.7, because of the relative homogeneity of the expertise data (due to the population selection), and a lack of significant results, we felt it might be best to simply remove this from the paper for clarity and readability.

2.4) Please better describe the test conditions. In section „Method: Experiment 1 – Paper vs. iPad“, „Procedure” you write: „Both the data presentation mode (paper or iPad) and data set order were counterbalanced.” So i assume you are varying 2 presentation modes and 2 test sets independently. So each participant could be allocated to one of 4 test conditions:

iPad Set1 – Paper Set2,
iPad Set2 – Paper Set1,
Paper Set1 – iPad Set2,
Paper Set2 – iPad Set1

(see also 3.7)

RESPONSE: We have clarified how the counterbalancing was conducted, and described the method of assignment. The clarification now reads “Both the data presentation mode (paper or iPad) and data set order were counterbalanced. Participants were assigned, in arrival order, to one of 4 data conditions: Paper first using data set1 followed by iPad using data set 2, Paper first using data set 2 followed by iPad using data set 1, iPad first using data set 1 followed by Paper using data set2, or iPad first using data set 2 followed by Paper using data set 1.”
2.5. Results
When presenting the results please report the values of both conditions (e.g. iPad and Paper) and not only the differences. In presenting only the differences there is a loss of information.

RESPONSE: We now present the values for each condition, for all of the experiments.

2.6) Method: Experiment 3 Procedure
“The exact procedure used in Experiment 1 was used to conduct Experiment 3.”
That’s not true. As with experiment two you should add: “with the exception that both mediums were assessed with the SUS after being used.”

RESPONSE: We have added this clarification.

3.1) The MDASI is an important element in the paper. Why not put it in the title or the keywords?

RESPONSE: We have added MDASI to the keyword list.

3.2) Section „Physician Interviews“:, paragraphs 2 and 3
“critical symptoms” and “criticality” are mentioned. Also in “Method: Experiment 1, Procedure”. Who determines what is critical?

RESPONSE: For the purposes of this experiment, criticality is defined arbitrarily, but in practice is defined by physician consensus so that a set of rules regarding when to contact a physician or patient can be established. We used the consensus values from the MDASI researchers, and specified that to the participants, as noted in the paper: Participants were also instructed that red indicated a critical item and that criticality was defined as any symptom item that earned a score of 7 or higher on the last visit or had an increase of 4 or more from the third to last visit. This is described in the text.
3.3) Section „Method: Experiment 1 – Paper vs. iPad“, „Materials“
As far as I understand there are no patient groups. So you can leave out the first part of this sentence: „Within a patient group, each patient’s data set was comprised of four completed MDASI surveys representing the patient’s MDASI six-month history“ and simply state: „Each patient’s data set was comprised of four completed MDASI surveys representing the patient’s MDASI six-month history.“ For more clarification you could add something like „So each participant has to face a test set of 80 completed MDASI surveys.” (see query 1.2)

RESPONSE: Agreed – we have modified the sentence as recommended and further clarification to the number of surveys was added in our 1.2 response.

3.4) Section „Method: Experiment 1 – Paper vs. iPad“, „Materials“ paragraph 2
I recommend that what is termed as „across-group information (Across-Group)“ should be renamed „between-patient information (Between-Patient)“ since I am still of the opinion that there are no patient groups in this study.
And as a better opposition to “between” you could rename what is termed as „across a single patient’s history (Across-Patient)“ into „within a single patient’s history (Within-Patient)“

RESPONSE: We agree that the longer names are more descriptive. However, they become quite cumbersome in the sections where we are discussing the groups in details – As noted elsewhere, we have expanded the paragraph clarifying the questions used, and the groups to which they belong. As part of that, we have clearly labeled each of the questions types so the reader has a ready reference.

3.5) The „six-question task list“:
You write „Five of the six questions were designed to judge either … or …“. You could be more informative when you number exactly how many questions were designed to assess information between the patients and how many were designed to assess information within a patient’s history.

RESPONSE: We have detailed the exact questions used, and described in detail what each was designed to assess.

3.6) Section „Method: Experiment 1 – Paper vs. iPad“, „Materials“ paragraph 2
You could briefly describe the SUS (comprising 10 items, scoring in the range of 0 to 100, …)

RESPONSE: An excellent clarification. We have briefly described the SUS in this location.
3.7) Can you describe how the testing order or the test conditions were allocated to the participants. How did you manage that „Both the data presentation mode (paper or iPad) and data set order” were counterbalanced? Were they randomized or consecutively alternated?

RESPONSE: We have clarified how the counterbalancing was conducted, and described the method of assignment. Please see 2.4 for the exact text used.

3.8) „Questionnaire scoring“
You write: „The remaining two questions had multiple parts, …“
You could be more informative if you tell the exact number of parts. e.g. „Of the remaining two questions one had 4 and one had 6 parts, …“

RESPONSE: The questions and their number of parts are now detailed in the sentence.

3.9) „Questionnaire scoring“
You write: „The remaining two questions had multiple parts, …“ You could be more informative if you tell what kind of questions had multiple parts ( across-group, across-patient or single-survey questions).

RESPONSE: The exact questions and their number of parts are now detailed in the sentence.

3.10) „Questionnaire scoring“
You write „the accuracy was calculated as (1/# of parts * number of correctly answered parts).“ As there is no real fraction line in the text please displace the closing bracket to be unambiguous: „the accuracy was calculated as (1/number of parts) * number of correctly answered parts.“ (and you could harmonize this term by either using “#” or „number“)

RESPONSE: We have harmonized the equation, as you suggested
3.11) „Questionnaire scoring“
„the accuracy was calculated as (1/# of parts) * number of correctly answered parts.“ This is more an idea than a revision: For a right answer of a simple question you get 1 point and for 3 correct answers of a 4-part question you get 0.75 points. Did you consider to score 1 point for each part of the multiple parts questions to give these questions a higher weight? If the multiple part questions are more mind demanding this would more adequately correspond to the performance. You can again (as you did) calculate a standardized average accuracy by dividing the sum through the maximum possible points and get an accuracy range from 0-1. But this is just an idea for I don’t know what the questions are asking for and you will have your reasons to score as you did.

RESPONSE: We have harmonized the equation, as you suggested in 3.10. Your suggestion for scoring is an interesting one. We used the current scoring method to keep in line with other work we have done, to make comparisons easier. In the future, however, we will consider this for new work.

3.12) Number of Experiments
You could describe your study as 2 Experiments. The first one (described as Exp1 and Exp3) with 27 participants (15 f, 12 m), mean age 28.7 (SD 6.7) where 2 subgroups could be distinguished (17 undergraduates and 10 medical professionals). If you would analyze your data this way you could first present an allover result (which would have more statistical power) and then you could take a look at the 2 subgroups and show that the overall accuracy and completion time effects are present in both groups and that the across group effect is present only in the undergraduates and you can juxtapose the SUS ratings of the two subgroups and so on.

RESPONSE: We had considered this way of analyzing the data, but in the end were persuaded by internal reviewers who commented that the presentation of the physician data as a separate study helped clarify that the results were still valid when tested with medical professionals, rather than undergraduates.

3.13) Results general
You could harmonize the results reporting of the 3 experiments. In some experiments some results are not reported that are reported in other experiments. E.g. the association between expertise and accuracy appeared only in experiment 1. You could try to use the same results pattern for all 3 experiments (with the exception of the missing SUS assessment for the paper in exp1)
RESPONSE: We have harmonized the results to the greatest degree possible, reporting the same results where possible, the notable exception being the SUS for experiment 1, where it was not collected for the paper task.

3.14) Experiment 1 Results Task completion time
   “… 2x (Q3) to 42x (Q1) longer response times”
   The abbreviations Q3 and Q1 appear in the paper without explanation. I suppose they designate question 3 and question 1, which is an useless information since the reader doesn’t know the wording of the questions. Either omit the “Q3” and “Q1” or list all 6 questions which would also help to clarify my queries 3.5) 3.8) 3.9) and 3.11).

RESPONSE: WE have eliminated the reference to individual questions through out, and have gone to just the descriptions of the types of questions groups. We also describe all the questions when first introduced and have attempted to address your concerns in 3.5, 3.8, 3.9 and 3.11.

3.15) Method: Experiment 2 – iPad vs. iPod
   You describe the merits of the smaller smartphones (also in the Discussion, section screen size you write “Using smartphones to view medical records …) but you use an iPod Touch which is not a smartphone. You could briefly justify the use of an iPod because it’s screen size resembles the screen size of contemporaneous smartphones.

RESPONSE: We have clarified that the ipod served a proxy for the iphone due to similarities in form

3.16) Discussion, paragraph 4
   “the mini-trend view (presented in the icon view)” For more redundancy you could once more refer to figure 1

RESPONSE: We agree that this is a good clarifying addition – we have added the reference to Figure 1 in this sentence.
3.17) Discussion, The effect of screen size
I recommend to remove the word “effect” and plainly state “Screen Size”.

RESPONSE: This has been corrected

3.18) References
To further improve the paper I would recommend to extent the related work section, to include more recent relevant literature from the field.

RESPONSE: Additional text has been added to include some of the most current research references in the area of electronic display of patient data.

3.19) Figures 2, 3 and 4
This three (actually 6) figures are not very informative. They contain 2 mean values (represented by the top line of a bar), the difference between them (which actually can be measured out of the two bars) and a SEM-bar which tells me something about the sample size but is inappropriate for the description of the data. I strongly recommend to either remove them without substitution because everything depicted can be said in the text or to use Plots that contain more information than presented in the text. Suggestions: A line graph with e.g. accuracy on the y-axis and iPad and Paper on the x-axis where each single value is represented and the 2 paired values of each participant are connected with a line so that the slope shows the individual change. Or dot plots (in which also each single value is represented, though not connected) or Box-and-Whisker Plots (which indicates at least 5 aggregated values)

RESPONSE: Several reviewers commented that these figures added little value to the presentation. We have removed them from the paper.

3.20 Method Experiment 1, Procedure
“… when the participant indicated that she had finished …” should read “that he had finished”

RESPONSE: The sentence has been re-written to remove the gender. It now reads “…when the participant indicated that the answer to the question was complete.”
3.21 Discussion, Screen Size
“ultra-portable” I find the prefix “ultra” inapplicable. It could be replaced by something like “easily portable and handy”

RESPONSE: Ultra-portable has been removed and the sentence now reads “Using a smartphones to view medical records is advantageous because the smartphone is quite portable.”

REVIEWER 2

1.1. As I have written in my introduction, my main concern is that you are comparing two different data sets - unstructured (in paper) and structured (in electronic format). This makes the value of the study less clear. I strongly suggest that the authors in much more clear statements argue about the scientific quality of such a comparison.

RESPONSE: It is true that we are examining the form of the paperwork that is currently used by the physician with a structured electronic format. We believe that this is still a valid comparison, since this is what is in use, and we have proposed a solution that has better efficiency and accuracy characteristics. Since many paper tasks may driven by the original form of the data, as it is collected, this research shows that this simple presentation format, while easy and expedient, may have consequences in the ability of physicians to utilize the data as intended. We felt there was a compromise that had to be made between internal validity (where we would examine unstructured paper/unstructured iPad) and external validity (which is the study we ran) and because we were interested in how the current status quo situation might be improved, we favored external validity. We have added a clarification about the importance of the comparison we did to the introduction.

1.2. Page 8, second paragraph: - “A six-question task list was used to gauge how accurately and quickly participants could answer questions about a given set”. Since the iPad data set already was installed on the iPad, it would be beneficial to know how time-consuming and complex that task was.

RESPONSE: In a related (but as of yet unreported) effort, we have been developing a method whereby patient data would be collected electronically using a telephone or web based interface and imported directly to a database which would then render the MDASI physician facing interface, so no additional data entry time, aside from that taken by the patient responding to the survey, would be required.
1.3. Page 10, Task Completion Time: - “As expected, we found that participants were significantly faster…”. Please indicate how this significance was measured. For example with $p =$ or similar figure.

RESPONSE: WE have added the results of the significance tests to this location.

1.4. Page 10, Participants: - “Forty undergraduates (25 female: 15 male)”. Please tell the reader if these were Medical undergraduates or if there also were other professions in this group.

RESPONSE: These participants were drawn from the general student population. We have added this clarification.

1.5. Page 13, Discussions. – “although the students tended to be faster and more accurate than the medical professionals for both paper and iPad presentations”. Please discuss possible reasons for this

RESPONSE: This difference was not statistically significant, so we have removed this sentence from the paragraph

2.1. In the Abstract, the term “technology history survey” is used with no explanation. As far as I know, this is not an established medical term, so please explain that to the reader

RESPONSE: This was a survey that assessed participants’ general familiarity with technology. However, as noted in the response to Reviewer 1, 1.7 decided that because of the relative homogeneity of the expertise data (due to the population selection), and a lack of significant results, we felt it might be best to simply remove this from the paper for clarity and readability.

2.2. Introduction, page 4 – “Though the MDASI contains pertinent treatment information, the act of obtaining information may be particularly time consuming and error prone”. I would like a reference for that assertion, to make it clear that this is not only a subjective feeling by the authors.

RESPONSE: This statement reflects our knowledge after the study was complete (where we do show that it is time consuming and error prone). Since this is in the introduction,
we have modified it to be more reflective of what we suspected before the study began. The new sentence reads “Though the MDASI contains pertinent treatment information, it might be difficult for a user of this data to synthesize the data from a large number of different paper sources for multiple patients or multiple visits for a single patient. For example…”

2.3. Introduction, page 5, just above the “Physician interviews” – “using medical professionals as participants.”. Please explain which medical professionals. Were they Physicians only or also Physician Assistants (PAs)?

RESPONSE: This is detailed in the Experiment 3 methods, where we identify the participants as 8 physicians and 2 registered nurses.

2.4. Page 7, last sentence – “two twenty-patient sets were…” – Were the sets of the same complexity, or not? Please make it more clear to the reader.

RESPONSE: Yes, the data sets were of similar complexity. We have clarified this by modifying our first description of the two data sets to read: “Since participants were instructed to answer patient-related data using both the MDASI paper survey and the iPad, two different twenty-patient data sets of similar complexity were prepared. A participant would use one dataset with one presentation format, and then another comparable dataset with the other presentation format.”

2.5. Page 11, top: - “As with the paper, the iPhone prototype…”. Do you mean iPhone or iPad?

RESPONSE: We meant to say iPod. In our interviews, the physicians talked about always having their phone available. We used the iPod Touch as a proxy for the iPhone, as they are nearly identical in form, but it was easier experimentally to use the iPod touch as a platform due to phone subscription issues. This has been changed in the text.

2.6. Figure 1. – Please provide the figure in color instead of in b/w to improve clarity.

RESPONSE: This figure has been updated to color
3.1. Page 9, Experiment 1 results: - Unexpectedly, we found an iPad advantage for only Across-Group items…” Please discuss this in more detail in the Discussions section.

RESPONSE: This is discussed in much greater detail on pages 14-15 in the Discussion section.

REVIEWER 3

1.1. The introduction and methods section should be reorganised in order to separate more clearly what is state of research and what has been done within this study. It may be helpful to include a study design section at the beginning of the methods to give an overview over the three experiments (this is currently part of the introduction). Separating background/state of research and methods more clearly is essential to increase readability of the manuscript.

RESPONSE: The flow and organization of the paper is meant to answer three questions: 1) is a medium scale electronic presentation (e.g. tablet PC) better than a paper presentation; 2) can this medium scale paper presentation be scaled to smaller, more portable devices (e.g Smartphones) and 3) do these results remain consistent when tested with a target medical professional user population. As you have suggested, we have added this clarification in the introduction to help frame the paper for the reader.

1.2. As the sample sizes in the experiments are very small we strongly recommend relying on non-parametric statistical tests. You should not use T-tests or ANOVAs for such small sample sizes. For such sample sizes it may be more meaningful to report medians and interquartile ranges (rather than means and SEMs).

RESPONSE: As you and other reviewers have suggested, we used non-parametric tests to evaluate the data throughout the paper. Medians and IQRs are reported, along with means and SDs to aid the reader in the interpretation of the data.

1.3. The results/methods section is not well-organised. It is uncommon to have several methods/results sections in one manuscript (the manuscript is currently structured according to experiment).

RESPONSE: The flow and organization of the paper is meant to answer three questions: 1) is a medium scale electronic presentation (e.g. tablet PC) better than a paper presentation; 2) can this medium scale paper presentation be scaled to smaller, more
portable devices (e.g. Smartphones) and 3) do these results remain consistent when tested with a target medical professional user population. We have added this clarification in the introduction to help frame the paper for the reader.

1.4. In the section “Physician interviews” there are no descriptive statistics given (e.g. how many physicians gave what answers concerning the five areas of interest? how many physicians emphasized criticality?).

RESPONSE: All of the physicians were in agreement with the importance of trend information, the importance of the pain measurement, the need for identification of critical items and the ability to map treatment onto individual MDASI scores. This has been clarified in the text.

1.5. You should not show patient names in your figure 1!!

RESPONSE: As noted in the methods section, the patient names were generated using a random name generator, set to mirror the diversity of patients found in Houston. These were NOT real names, and this is NOT actual patient data, but data meant to be similar to real patient data in its characteristics.

1.6. The description of the SUS scores should rather be given in the methods section than in the discussion.

RESPONSE: The descriptions of the SUS scores are given to help explain our understanding of why the usability metrics came out as they did. We cite several benchmark studies to show how these results relate to the larger usability scores reported in the literature, and to give semantic reference anchors to help the reader interpret the scores. We have also added a description of the SUS in the methods section.

2.1. It may be more informative for the reader to convert the figures 2-4 to a table

RESPONSE: As other reviewers have suggested, figures 2-4 did not add substantively to the readers understanding of the data and have been removed.
2.2. The title of experiment 3 refers to iPod, I guess you rather meant iPad like in the text below.

RESPONSE: Indeed, this was supposed to say iPad. It has been corrected.