Reviewer’s report

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

Version: 2 Date: 8 February 2013

Reviewer: Gerold Schwantzer

Reviewer’s report:

In my opinion this paper is relevant, well structured, well written and easy to read.

The design of the experiments is conclusive and founded on reasonable preparatory work.

However some details are missing and the statistics is inappropriate.

- Major Compulsory Revisions

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?

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.

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.

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.

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.

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.

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).

- Minor Essential Revisions
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.

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 …“

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.

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)

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.

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.”

- Discretionary Revisions

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

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?

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)

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 patients’ history (Across-Patient)“ into „within a single patients’ history (Within-Patient)“

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 patients history.

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, …)
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?

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, …“

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).

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“)

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.

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.

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)

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).

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.

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

3.17) Discussion, The effect of screen size
I recommend to remove the word “effect” and plainly state “Screen Size”.

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.

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
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)

"Minor issues not for publication"

Method Experiment 1, Procedure

“… when the participant indicated that she had finished …” should read “that he had finished”

Discussion, Screen Size

“ultra-portable” I find the prefix “ultra” inapplicable. It could be replaced by something like “easily portable and handy”

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

**Quality of written English:** Acceptable

**Statistical review:** No, the manuscript does not need to be seen by a statistician.

**Declaration of competing interests:**

I declare that I have no competing interests