category	reviews	revision
Participants	As such the results mostly reflect what people were already interested in doing and tracking fitness data is expected from such a tool. Other groups of people may have different feelings and expectations what such an app should offer in order to engage them in fitness activities. (about generality) (R1)	We added a discussion of this as a study limitation in section 6.3
	In terms of generality, a potential obstacle might be the rough characterization of the participants. It is not clear how participants were recruited, what their lifestyles look like (profession, kids, self-assessed fitness level, etc.), how long they have been using Fitbit, etc. There is a number of demographic factors that could influence people's attitudes and engagement. I don't think this would influence much the findings related to the on-calendar visualization, but it certainly has impact on the validity and generality of the feedback model, since it was empirically derived. (R3)	We noted this in the study limitations in section 6.3
	With regard to selecting participants I was also wondering how they were selected (e.g., randomly)? (R1)	We sent out invitations to recruit participants on campus and through social networks. The screening criteria are in Section 4.1
Quantitative data report	I assume an independent samples t-test but this is not explicitly stated. Also, as PA values were measured weekly were those averaged for the statistical test? (R1)	We added more details about the t-test, and report degrees of freedom, exact p value and Cohen's d in Section 5.1.
	It is not clear what procedure was used for statistical test, the nature of the response variable, and the parameters of the test. For ANOVA, for example, report F-value, degrees of freedom and the exact p-value. (R3)	
	I assume the reported p-value should read p = .53, as < .53 is not very meaningful. (R1)	
	I would have liked to see more quantitative data on the use of the application, especially a time progression of the use of the tool. The authors only present the overall use, but a use per week would help show habituation, or real use. (R2)	we added a chart of time progress of the usage in Figure 2 (bottom)
	Section 5.1: It is also unclear how PA tended to increase more for the experimental group. Is this observation based on the outlier in the experimental group? Usually, to back up such a statement, one would model the data with a linear model and report the slope estimate for the group effect. (R3)	We understand the reason the reviewer suggested a linear model. However, in this case, the independent variable is categorical, so the slope is simply the difference between the mean values of the two groups. It might not be very informative, but we added cohen's d to reflect the effect size in section 5.1.
	Figure 2: This chart does not appropriately represent the distribution of scores because it is hard to judge how many observations overlap (some vertical lines seem to merge, making it hard to assess density). An easy solution would be using circles, as in Liu and Heer (2014). (R3)	We changed the vertical lines into circles in this chart (now Figure 2 top).

category	reviews	revision
	The overall statistical reporting of the study feels a bit lacking to me. I would love to have seen more information about which participants used the tool and how much. This information would provide very useful context when interpreting the more qualitative discussion that follows the current statistical reporting is sparse enough that it is difficult to gauge how heavily the participants actually used the tool, how their usage varied over the course of the deployment, and how the activity patterns of various participants compared – all information that would add extremely useful context to the overall discussion. The statistical reporting that is provided (e.g. in section 5.1) is limited and mostly involves reporting simple counts and p-value results for null-hypothesis tests. Using graphical reporting techniques to highlight the distribution of usage and activity, and using plots with confidence intervals rather than tersely-described statistical tests would make the discussion of the study results much more intelligible and convincing. For more information on modern best practices in statistical reporting in HCI/Vis, I suggest the authors consider http://www.aviz.fr/badstats. (R4)	We added a chart to reveal usage over time (Figure 2 bottom). However, we did not go into greater depth than this because our main focus was on qualitative analysis; observing behavior differences was not our main goal (see argument for this approach in Section2.3). We also revised the t-test reporting in section 5.1 and added the cohen's d to reflect the effect size.
Qualitative data report	Section 5.3: Sometimes it is clear how the statements separate into experimental group and control group, sometimes not. This could be made more consistent, e.g., by using something like (9/9 and 10/10) (R1)	We revised the report of counts as the reviewer suggested in section 5.4 (page 5). In the rest of the sections we used the label V and C to identify the groups.
	I was missing some sort of summary of the most important findings at the end of the paper. The paper discusses a lot of results, as such, digesting the most important results and possible implications for the design would really be very helpful (R1)	We added a brief summary of findings and design implications in the conclusion.
About the model	One of the strongly-emphasized outcomes of the paper is the authors'model of the behaviour feedback process. However, I found the discussion of the model somewhat unsatisfying, and the model itself so abstract that its unclear to me how other researchers might apply it. In their discussion at the end of the paper, I would appreciate a more concrete discussion of how this conceptual model might be useful going forward. What kinds of questions or possible designs does this suggest? How could it be used to examine other kinds of persuasive or reflective systems? (R4)	We added a brief discussion of design implications in section 6.2 (paragraph 2), with respect to the feedback model.
clarification	In Section 6.3 it is mentioned that further studies should investigate usage on mobile devices but on p6. (Section 5.6) it is stated that some participants already primarily used their phones. Please clarify.(R1)	In the study, we saw people already use apps on their phones, provided by Fitbit or other companies. However, the current version of our application was primarily designed for desktop use. We suspect people might have interacted with it differently if it had been customized for mobile devices. We added clarification on this point in the Limitations section.
	Up until Section 6.3 I had the impression that the application was integrated into the Google calendar. Perhaps this could already be made more clear in Section 3.2(R1)	Because of limitations imposed by Google, we copied the look and features of Google calendar, and our application is on a different online portal rather than Google itself. We added the clarification in section 3.2.

category	reviews	revision
	I have one piece of criticism regarding the conclusions about preferred visualization settings (Figure 3) for the on-calendar visualization. It is unclear what are the default settings of the tool, and whether they could have influenced participants' preferences. Another source of bias is instruction: what settings were on when participants received instruction on how to use the tool? A possible way to mitigate this threat would be providing the tools with different (random or balanced) settings. Since this cannot be done, I suggest acknowledging the possibility of bias and describing the default settings. (R3)	When the application was introduced, participants were asked to try to explore all possible settings. The application was implemented to remember the customized settings, so the bias of default visualization settings was minimized. We added clarification about this point in section 5.2.
	On Fitbit: The paper is largely based on Fitbit data and its feedback tools (baseline condition in the field study), but they are never described. This is a problem to the reproducibility and archival of the user study. If Fitbit goes out of business next month, ten years from now it could be hard to get detailed information about Fitbit data and tools. In addition, Fitbit's applications and data will evolve, so the paper needs to contain a "snapshot" of what these components were like at the time of publication. It is clear to methe data wasn't used for measuring physical activity, but it was used on the on-calendar visualization (R3)	We agree, good point. We added a screenshot to supplementary materials due to the page limit of the paper itself.
	I would suggest the authors change the label “experiment group” to “visualization group” since the word “experiment” really relates to the fact that you are running an experiment, so technically the control group is also part of this experiment. It was confusing at times, especially at the beginning. (R2)	Changed as suggested.
	Unfortunately, the paper fails to track and account for the Fitbit app entirely; the users were allowed to continue using it, even in the control group. I think this is something that should have been either (a) limited, or (b) explicitly tracked. (R5)	The purpose of our control group was to baseline the Fitbit app use. Unfortunately, frequency of using the Fitbit app would not have been feasible to track except through unreliable self-reports. We added a discussion of this point in the study limitations (section 6.3).
Writing	in terms of presentation the text itself is sometimes quite densely packed. For example, the explanation of the categories could be layouted as some sort of list. This would help if the reader wants to refer to the description of the categories again.	This is a good suggestion but was unfortunately not possible due to the page limit.
	While the writing is generally quite good, the paper includes a large number of passages written in passive voice. This obscures agency particularly in the discussion of the study, and makes it less clear who did what. I encourage the authors to check for and rewrite these passages in their revision. (R4)	We re-worded the passive tone in the discussion.
Other	In a way, given the paper is written, a lot of the findings here are actually reflections of how people use personal fitness trackers rather than the integrated time-series visualization proposed as a contribution in this work. On the one hand, this means that the paper is not fully on target and relevant for the topic at hand. On the other hand, the fact that this work has enabled this kind of "secondary findings" should not be counted against it; on the contrary, it shows the premise of calendar visualization and its utility. (R5)	We are aware of the difficulties and challenges of evaluating ambient applications, especially those with a non-persuasive perspective. We hope the study will provide designers a different view of designing visualizations for feedback purposes, in which the on-going effect and people's existing information use habits need to be considered. We added design implications in Section 6.2 and revised the study limitations in section 6.3.