NYU

Junior Garcia <jfg388@nyu.edu>

CHI2024 - Reviews for paper #5486 - not accepted

1 message

Phoebe Toups Dugas, Irina Shklovski, and Max Wilson <chi24a@precisionconference.com> Fri, Jan 19, 2024 at 3:01 PM

Reply-To: papers@chi2024.acm.org
To: Junior Garcia Ayala <jfg388@nyu.edu>

Dear Junior Garcia Ayala,

As a follow-up from our last notification email, we are now sharing the final reviews of your paper #5486, CREST: Mediating Collaborative Search and Agreement for Group Property Bookings.

Although your submission was not accepted after revision, we hope that revising the paper made it stronger, and that the feedback and development is useful for improving your paper further for future submissions to CHI2025 and elsewhere.

Please note - reviewers were invited to: change recommendations if relevant, add re-review detail if needed, and contribute to discussions with the Associate Chairs. Consequently, reviewers did not necessarily enter text into the re-review box. You do not need to email us if a re-review box is empty - this does not affect the decision of your paper, which was made and agreed upon during a PC Meeting between senior reviewers.

Best regards,

Phoebe Toups Dugas, Irina Shklovski, and Max Wilson
CHI 2024 Papers Chairs

1AC review (reviewer 4)

Expertise

Knowledgeable

Originality (Round 1)

Medium originality

Originality (Round 2)

Medium originality

Significance (Round 1)

Low significance

Significance (Round 2)

Low significance

Research Quality (Round 1)

Low research quality

Research Quality (Round 2)

Low research quality

Recommendation (Round 1)

I recommend Reject.

Recommendation (Round 2)

I recommend Reject

1AC: The Meta-Review

The reviewers all compliment a well thought through paper that systematically presents a set of challenges, translates them into design principles, realises those principles in a functional and feature rich system, and evaluates that system through a user study.
The reviewers also comment on the clear description of goals and motivations (R1, R2) and a generally well-written paper (R1).

There seems to be a general consensus that CREST is an impressive system where a lot of thought has gone into its design.

However, the reviewers also point out several shortcomings of the paper.
All reviewers struggle with seeing what the core research contribution of the paper is, and what general takeaways can be derived from the design and the study.
What makes it difficult to assess is, on the one hand, that both the design goals and the task analysis seem to be mostly anecdotally based and not rooted in empirical studies or the literature. R1 points out that without some form of evaluation of these, it makes it difficult to assess their validity. On the other hand, both R1 and R2 have reservations about the evaluation. R1 is not convinced by the comparison to a baseline that is just a cut down version of CREST and misses more qualitative insights into how the collaboration using CREST unfolded.
R2 also found the qualitative part of the evaluation to be lacking, and questions the bias a reward to the participant that had their preferences satisfied the most might introduce. R3 is not convinced by the statistical analysis given the low number of groups participating in the study.
While the reviewers general find the paper well written, they also highlight numerous structural issues. Particularly, R3 points out that it would be beneficial to have the related work more conventionally after the introduction, which could also potentially help positition the design principles and task analysis more strongly.

Overall, the paper presents a thoughtful system design, but the research contribution as the paper is presented is unclear and limited. The rewriting and restructuring needed to address the reviewers concerns are substantial, and beyond would we can expect for a five week revision. I therefore recommend rejecting the paper.

1AC: The Meta-Review (Round 2)

All reviewers appreciate the effort the authors have put into revising the paper. However, they also all have concerns that makes me reluctant to recommend accepting the paper.
The changes to the paper are quite significant and 2AC points out that it is difficult to couple the summary of changes in relation to the themes of the reviews.
R1 still struggle with identifying a significant contribution and R2 finds that the added section 3.1 introduces a new weakness of the paper in terms of a lack of rigour in reporting on qualitative findings. This is particularly an issue given the paper's reframed focus on the qualitative findings.

The reviewers were on the fence with the paper to begin with, and there are still several unresolved substantial issues in the paper that leads me to recommend rejecting the paper.

2AC review (reviewer 3)

Expertise

Passing Knowledge

Originality (Round 1)

Medium originality

Originality (Round 2)

Medium originality

Significance (Round 1)

High significance

Significance (Round 2)

Medium significance

Research Quality (Round 1)

High research quality

Research Quality (Round 2)

High research quality

Contribution Compared to Length (Round 1)

The paper length was commensurate with its contribution.

Contribution Compared to Length (Round 2)

The paper length was commensurate with its contribution.

Figure Descriptions

Recommendation (Round 1)

I recommend Revise and Resubmit.

Recommendation (Round 2)

I recommend Reject

Review (Round 1)

This paper introduces and evaluates a system to support all the stages of real estate booking (i.e. search, discuss, and agree) as a full pipeline, building on previous work which usually cover a subset of these stages.
I appreciate the thorough effort of the authors in laying out the existing challenges (e.g. discontinuity, disconnect, etc.) to build design principles in addressing those challenges. The authors also did a good job of transferring these design principles to a working prototype to evaluate. Below, I describe a few points to improve the work:

- The structure of the system: While the authors did a good job in describing the challenges, and design principles, the transition to the system description was not this smooth. The section “2.3 Design Principles & Implementation” is a large chunk of the text that doesn’t have much organization, and the reader gets introduced to different parts of the system like it’s already discussed before, while not. It’d be helpful to have a chart/overview description of different elements of the system (like different panels), maybe it can be done by revising the Table 1 to add those UI elements.

- Related work: While I understand that in some fields the related work goes last in a paper, this is not a norm for the CHI community, particularly in the case of this paper that when it jumps directly to the Design, it makes it difficult to see what this new tool has that the previous work didn’t; so I recommend the authors to bring related work to the beginning of the paper, after introduction.
Also, in related work, I don’t see a discussion about the large corpus of work in collaborative crowdsourcing, and how different mechanisms have been designed to facilitate people working together in different processes.

- Statistical analysis: There are several places in the paper that statistical analysis looks like a stretch, particularly that there's only 7 groups. Has a power analysis been done to see if this number is enough? So I advise the authors to carefully revisit each of their statistical analysis, and remove those that are not valid for this size of data.

- Misc.
The small figures in page 16 (the ones with the blue/red dots) are not clear. If it follows Fig 6 labels (e.g. red dots are "search", etc), make that clear.

Fig 6: the coloring isn't the best way to show different activities (the discuss and agree are similar in color, and also when the paper is printed in black/white, or for color-blind people, there's not enough distinction between the colors, using patterns/different shapes can help)

Re-review

I appreciate the authors' effort in addressing the comments, however, it was quite difficult to navigate the many reviews and changes applied in the previous rounds of this paper, so I mainly focused on the changes the authors reported for this conference. The summary of changes was still difficult to navigate through as it was not organized based on themes of the reviews, but based on what I followed, there are a few comments that have not been addressed:

- The structure of the system: I didn't see any structural change or revision for the "Design Principles & Implementation" section to make navigation easier (I see that the authors added the references of figure to Table 2, but it's not enough as the text itself needs some reorganization and clarification)

- I couldn't find the added literature and discussion about collaborative crowdsourcing in the marked document.

- For statistical analysis, I understand that this is qualitative work, and it's ok to therefore focus on quantitative results, but my main concern, as I mentioned, is that right now there's a report of some statistical analysis that, as the authors themselves mentioned, is far enough based on the power analysis. I appreciate the authors adding that to the limitation section, but this is not enough for addressing this. Rather, I expected to see the unnecessary/unjustified statistical analysis removed. Right now, there are still a lot of results that just make the final interpretation muddy.

---

reviewer review (reviewer 1)

Expertise

Passing Knowledge

Originality (Round 1)

Low originality

Originality (Round 2)

Low originality

Significance (Round 1)

Low significance

Significance (Round 2)

Low significance

Research Quality (Round 1)

Low research quality

Research Quality (Round 2)

Low research quality

Contribution Compared to Length (Round 1)

The paper was too long in addressing its claimed contribution.

Contribution Compared to Length (Round 2)

The paper was too long in addressing its claimed contribution.

Figure Descriptions

The figure descriptions are adequate and follow the accessibility guidelines.

Recommendation (Round 1)

I can go with either Reject or Revise and Resubmit.

Recommendation (Round 2)

I recommend Reject

Review (Round 1)

Contribution

The primary contribution of this work is the presentation of CREST, a tool for collaborative renting of apartments/houses/etc. Design rationale and a user-study based evaluation of the tool are presented.

Review

This submission presents the thoughtful design of a system, CREST, to facillitate group collaboration in renting properties (e.g. apartments, houses, etc.), such as for group vacation rentals or roommate contexts. It is primarily a case study of a system. The submission is well-written and clearly describes the design goals and motivations, the system behaviors, and intended use. The user study aims to demonstrate the tool's superiority to a fictional Baseline (not presented) that integrates some subset of the CREST functionality. The underlying design conceit in this work is that all of the process should be handled within a single tool, obviating the need for out-of-band communication with the aim of better fostering collaboration and group satisfaction in the resulting property booking.

As such, I am hard pressed to identify the research contribution that this submission aims to make. It reads in many ways as an advertisement for a specific system, and does a good job of presenting the design goals and rationale in it. For this reason, I characterize it as a case study. It is thoughtfully designed and shows some interesting design decisions and rationale.

However, it is difficult to discern what specific research contributions it makes. It is not clear where the design goals come from. They are clearly stated and articulated, and supported with references to the literature. But did they come from a careful analysis of the literature? From observations or interviews with users? In the absence of a clear description of their provenance, it is difficult to argue that this specific set of design goals makes for a research contribution on their own. Why these goals and not some others? Are these reasonable goals to strive for? Again, there is some support through references to the literature, but a more documented and systematic approach to their elaboration would be necessary to be able to argue that this articulation of design goals makes a meaningful contribution. Similarly, some sort of evaluation (or justification) of their appropriateness/sufficiency/etc would be necessary.

Similarly, the task analysis focuses on three main phases: search, discussion, and final agreement. This task analysis is interesting, but lacks backing. Did it come

from a careful literature review, from observations of users, interviews with participants in a formative study, or thoughtful consideration? It would seem to be the latter, but the provenance is not clearly described. Drawing from an extensive and careful literature review or from empirical methods would provide supporting evidence in support of this task decomposition. Or, should they come from thoughtful consideration, some sort of supporting evaluation that demonstrates that they are a useful lens would strengthen their weight. As presented in the submission, such backing is lacking.

The proposed system itself is interesting and clearly defined. However, the approach seems to argue that one of the main contributions of its design is the holistic nature of encompassing all of the collaborative property booking process in the task analysis, from search to discussion to final agreement. The submission seems to take the perspective that that is in and of itself an important goal. I am not sure I agree; would not more specialized tools for each phase—as long as the barriers between them are not too great—be better suited? I'm sure that it really depends, but the submission does not seem to provide supporting evidence one way or the other.

The remaining part of the submission is the evaluation. The evaluation is clearly described and appears well-executed. The quantitative analysis of the study results is thoughtful, and takes care to couch and nuance its interpretation. However, I am not convinced by the design itself: a) comparing CREST to a Baseline system implemented just for this study using a subset of the CREST components, and b) focusing primarily on quantitative metrics and analyses.

I am not convinced that a comparative study is the right evaluation to conduct. Given that the specific research contributions are not well-articulated, and thus the objectives and supporting metrics ill-motivated, it is difficult to determine whether these specific measures are particularly meaningful. Moreover, the baseline condition is not clearly described and presented, so it is difficult to evaluate just how meaningful the comparison is. Since one of the articulated insights seems to be that a holistic approach is better suited, I would have expected to see a study attempt to probe that approach, such as by providing a collection of well-suited tools for the various phases of exploration but with more exposed seems between the phases. But that would require a different study design and would require that be a specific study objective. Instead, *Baseline* seems to be invented just to give CREST *something* for comparison. In any case, this type of a study would require a more thorough description of Baseline for it to serve as a benchmark.

Instead, for this type of contribution I would have expected to see a more qualitative approach. (That is not to say that the quantitative analysis is not useful or well done; on the contrary!) It seems to me that what is more meaningful is less the quantitative use of the tool and how long people took to reach agreement but rather the qualitative nature of those exchanges. For example, the metrics suggest that Baseline users reached agreement later in their normalized exchanges, but the nature of their divergences, their agreements, and their explorations seem more pertinent. Did Baseline users explore more? Did they use the same strategies when collaborating as with the CREST group? These kinds of observations would seem more useful for such a context where specific quantified epistemological objectives are less well-defined—and are more robust to the issues of small sample size that the paper recognizes. I would have found such a qualitative evaluation more compelling in giving a richer picture of the nature of collaboration in this kind of a context, rather then a details focus on use metrics of interface components specific to this particular tool from which it is more difficult to draw conclusions.

For all of these reasons, I recommend that this submission undergo substantial revisions. In particular, I would like to see a more complete and clear statement of the contributions this work makes. In its present form, the contribution seems to be a tool. But what are the specific underlying goals of that tool? If it is to probe the role a holistic, end-to-end tool in collaborative decision-making (using property bookings as a lens), then those goals need to be clearly stated and the design probe or evaluation study more clearly articulated around those goals. If it is to probe the role of visualization of group dynamics in the collective decision-making process, then that needs to be clearly stated and the design probe or evaluation study more clearly articulated around *those* goals. If it is some combination of the above, then *that* needs to be clearly stated and

the evaluation constructed in a way that probes how they contribute to the overall experience.

This submission presents a nice system design, is well-written, and thoughtfully articulated. For the reasons above, however, I would argue that it needs a substantial refactoring to focus on specific research objectives and to perform a more qualitative evaluation that characterizes its use.

Nits & Misc Comments
---

p1, abstract: The mention of AirBnb and “group booking exercise” in the abstract suggests a framing of this work in terms of short-term vacation rentals, but then goes on to describe mediating bots and house rules, which don’t really seem to apply for such a context. It might make sense to more clearly articulate this more generally first as for group booking exercises, whether they be short-term (AirBnb-like) or long-term (roommates).

p1, l.33: “or other” is awkward…. “reunite” doesn’t seem like the right word here; it suggests people who knew each other well getting back together again.

p1, l. 36/37: “are capitalizing on market demand” sounds awfully startup market-speak and like an ad.

p2, l. 83: “technology and research gap” : I can see the need here, but it is not clear just what _are_ the technology and research gaps that are being addressed. This comes back to the comments above about articulating the research contributions of this work, such as by explicitly stating the technology and research gaps this work aims to bridge.

p2., ll.86–89: What evidence supports this argument? Neither the motivating text nor the user study provide empirical support for this claim, and the study design does not seem to probe this.

p2, task analysis: I would recommend explicitly drawing on or connecting to the information foraging literature (e.g. Pirolli, CHI ’09, 10.1145/1518701.1518795).

p4, l.164: Is it _really_ all that useful to support these in a single tool? This seems to articulate that as a given but I’m skeptical that a single tool wouldn’t be better than three purpose-built specialized tools optimized for each sub-task so long as they can talk to each other seamlessly enough

p4, ll.193–198: See the Pirolli & Card sensemaking model available at https://urldefense.proofpoint.com/v2/url?u=https-3A__www&d=DwIDaQ&c=slrrB7dE8n7gBJbeO0g-IQ&r=LFgT5eHJjh739rLb4hZAVQ&m=G5877F3itS93Fylp9aQ4jDz0D6lOSrIz9MUTcd2MbAGoSdNhmnPa2zylUXK2z6FP&s=uXe409X2bzZacDMmvW8EQG4DtYv2pj1N0cvH6BYj-Zc&e= .
e-education.psu.edu/geog885/sites/www.e-education.psu.edu.geog885/files/geog885q/file/Lesson_02/Sense_Making_206_Camera_Ready_Paper.pdf). This seems to conflate what Pirolli & Card describe as the “shoebox” and “evidence file” phases. (NB: this point is briefly raised in the study discussion but I would have liked to see some more consideration.)

p5, l.233: “deprive group members of autonomy” Needs supporting evidence/citation.

p5, l.241: “we aim to safeguard our users by protecting them from possible decision-manipulation scenarios” … again, needs supporting evidence to show a) how this aim is satisfied and b) whether it is successful

p6, table 1: I’m not sure how to read this. For example, what does the “unified system” row mean?

p7, fig 2: It should probably be mentioned that this screenshot depicts the interface for a short-term weeklong rental. I was a bit horrified at Liz’ contract clause that she should get exclusive use of the living room every Friday night.

p8, ll.397–398: Needs support.

p9, l.438: MeetingMediator does no such thing. It demonstrates a different visualization of a different set of engagement metrics using different stimuli in

a completely different context.

p12, l.601: Baseline needs more detail. Even just detailing the specific features would be better—but still insufficient. Design matters. That's the whole argument behind this design study. Without understanding the design and interaction, there is no way to understand *Baseline* as, er, a baseline.

p13, l.630: Participants were not queried based on their prior experience with the domain task (roommate rentals)?

p13, l.641: I find the avatar study design interesting—a nice, elegant way to deal with the issue.

p13, ll.654–655: "The anonymity of the participants ensured that the tool itself was the only communication medium amongst the team members." Is this really realistic for a benchmark comparison? In existing tools, people use out-of-band communication all the time! This seems like it would be hobbling the benchmark.

p15, figure 6: I really like this figure. This evidence could be useful in part to support a different research contribution that aims to provide empirical support for the task decomposition and more deeply explore the nature of the search/discuss/agree phases.

p16: figures: are these normalize relative to the stream of event sequences in the system? Since the Baseline does not include the contract phase events, wouldn't that mean that agreement would be structurally later in this stream for the Baseline condition than for the CREST condition?

p20, ll.991–993: Needs supporting evidence.

p20, §3.3: I would argue that the biggest study limitation is that it is primarily quantitative in nature rather than qualitative to reveal insight into what participants were really thinking as the encountered the stystem.

Re-review

I thank the authors for their detailed revisions and comments. However, while some of my points have been addressed in revisions, they do not seem to address the fundamental weaknesses of this work: this is a system contribution that does not seem to make a significant research contribution. The tweaks to the study analysis are an improvement, but do not fundamentally change the overall experimental design. As such, I find myself maintaining a negative recommendation.

---

reviewer review (reviewer 2)

Expertise

Passing Knowledge

Originality (Round 1)

Low originality

Originality (Round 2)

Low originality

Significance (Round 1)

Low significance

Significance (Round 2)

Low significance

Research Quality (Round 1)

Medium research quality

Research Quality (Round 2)

Medium research quality

Contribution Compared to Length (Round 1)

The paper length was commensurate with its contribution.

Contribution Compared to Length (Round 2)

The paper length was commensurate with its contribution.

Figure Descriptions

The figure descriptions are adequate and follow the accessibility guidelines.

Recommendation (Round 1)

I can go with either Reject or Revise and Resubmit.

Recommendation (Round 2)

I recommend Reject

Review (Round 1)

Summary
===
The paper describes a system called CREST to support collaborative property search; and a user study (with qualitative and quantitative findings) to evaluate its main design principles.

The user study had 2 conditions comparing CREST with a baseline tool (21 participants per condition). Each participant was given a persona, described as a personal-shopper of an avatar with specific requirements for their home. Participants were tasked to come to an agreement with their group by signing a contract. They collaborated asynchronously and remotely, in groups of 3, for up to 5 days.They were paid 14 USD, but a bonus was given for “the participant in the group that best represented their avatar’s requirements in the final selected property”.

One of the main hypotheses behind this paper is that a mediating agent (a chat bot provided in CREST) can help groups reach satisfying agreements.

All groups in both conditions eventually reached agreement (except 1 group from baseline condition) but differences in favour of CREST were only marginally statistically significant.

Strengths
===
+ Clear description of the problem domain and the motivation behind this work. The case study, even though very specific to property search, is interesting.
+ A functional system, with many features that seem to be implemented.
+ The link between tasks, challenges and design principles is nicely summarised in table 1. Although I find the motivation for the tasks not as strong (see comments below)
+ I like the categorisation of the mediator roles and how this work has been used to design the chat bot and in the user evaluation.
+ Overall, I find the user study method sound with some caveats (see my comments below).

Weaknesses
===
The analysis tasks & Challenges:

=
are not clearly motivated, and the same goes for the challenges. Although they seem reasonable, I am missing more motivation or justifications to strengthen the design principles. The same applies to the process of a collaborative search, which seems to be motivated mostly by the author's own experience or views: “While these tasks may seem sequential, we find that users often engage in them in no specific order, simultaneously and repeatedly.”.

Study task potential bias
=
From earlier discussions in the paper the notion of ‘fairness’ seems to be important to the design of a collaborative search tool (e.g., in sub-task 4 and principle 3). I do not see this notion discussed much in later parts of the paper. In the user study, a monetary reward was given to “The participant in the group that best represented their avatar’s requirements in the final selected property”. There is a risk that this bonus could have driven participants more towards personal gains than group agreement or satisfaction. This might also explain the observation described in section 3..2.3: “We found that in one of the two groups, a single user dominated the task and got what they wanted (hence reporting a score of 5 in terms of satisfaction) while ignoring the needs and wants of the other group members, who reported much lower satisfaction scores (2 and 3)”. Could this observation be due to personal/personality differences or due to the study bonus ?.

Missing details about the qualitative coding of the chat messages
=
This is especially for the ‘agree’ events. This was described as follows: “if a message mentioned a specific property, it was most likely seeking agreement on it and was often labeled ‘agree’, other messages were often labeled ‘discuss’". I am not sure why mentioning the word 'property' in a chat discussion results in coding this as agreement. An example would have been helpful (as well the supplementary material related to the coding, i.e. the codebook). This is important to clarify as one of the conclusions/results is that CREST supports more early / agreements than the baseline.

Takeaway messages and implications for the broader CHI community
=
With the main results only marginally statistically significant (RQ1 and RQ2 in particular) and the lack of depth of (or description of) the qualitative analysis, the takeaway messages to the broader CHI community are hard to pin down. I am left wondering about the possible recommendations for designing collaborative search tools, with regards to for example the design of chat bots for this context, evaluation / collaboration metrics and visualizations etc. Also, section 4 is a mix of related work and the discussion of the user study results, which makes it hard to find the main takeaway messages from this study.

Readability of the paper
=
Section 2.3 is rather long, mixing tasks, challenges and design principles together with the actual implementation and UI widgets. I wonder if a specific section dedicated to the user interface/system alone would help improve the readability of this section, and to get a better sense of how the collaborative search works within CREST. Again on the readability of the paper: table 1 has not been properly explained in the text. What are the main messages here? Is it that the chat bot supports all tasks and addresses all challenges? Has this been demonstrated in the user study?

There are also various places in the paper that are not clear and can be explained better especially with regards to the various metrics such as the difference between relevance and rating. How is the influence metric calculated? I also find the description of the system architecture of CREST (figure 5) not very informative. The designs of the avatar descriptions was not discussed: do all the criteria/requirements have the same importance? How were the personas designed? Were there specific characteristics that the authors had in mind when writing them, such as the number of criteria in conflict, “amount of conflict”? Were there variations in the results between the three personas?

Minor issues and typos
=
- Page 8 “the degree of engagement of the group members in the collaborative task through a visualization of user activity, Collabo-ratio (Figure 1),” ⇒ figure 3.
- Unsure how to read the thumbnail images in section 3: Do the bins correspond to the three tasks?
- Typos in lines 953, 981

Justification
===
Overall I find the paper interesting, I like the user study, and I am impressed by what looks like a functional system that has been implemented. However, my understanding is that the main contributions of this paper are in design principles and the evaluation of these principles through the CREST system. Overall the results are not very strong and the takeaway messages are not clear either, hence my somewhat low overall rating.

Re-review

I thank the authors for submitting the rebuttal and for making the new changes and improvements to the paper. My initial concerns with the previous version of the paper were related to:
1. Weakly motivated analysis tasks, challenges and design principles
2. Study task potential bias wrt to bonus
3. Missing details about the qualitative coding of the chat messages
4. Takeaway messages and implications for the broader CHI community
5. Readability of the paper

The authors successfully addressed my concerns with regards to 2 and 3, and partially for 5.

However, I still have reservations with regards to issues 1 and 4. I don’t think these have been adequately addressed.

In particular, the new section 3.1 describes an interview study to motivate the tasks, challenges and design guidelines. In principle this is a good addition, however, the new section brought new weaknesses to the paper. Besides the low number of groups (only 2, especially that in this case the authors are not dealing with domain experts and thus could have more easily added more groups), the main issue for me is the lack of rigor in reporting how the interview study was conducted, how the collected observations were analysed and how the findings back up the design guidelines. For me this is still a weak spot in the paper. Examples of lack of rigor in reporting include: what was the exact task given to the participants? How long were the interviews? Was there an interview guide? Was googledoc the collaborative platform of the participants or the authors? How were the observations gathered during the interviews? etc

As a result, Section 3.1 unfortunately does not tell a convincing story from the interview/observations, to the take-aways and tasks. This is partly due to the unsystematic reporting and analysis, and also the small number of groups.

The authors also state that the identification of tasks comes from a study of the available literature [10, 18, 28, 33, 39]. I am not sure whether the takeaways from this study were clearly described (or summarised) somewhere in the paper (e.g., discussing gaps in the literature, and contrasting this body of work with findings from the interview).

Besides - the paper claims a new focus on qualitative findings. I like this new focus/direction but I think the current reporting on the qualitative findings can be improved, going deeper into the analysis of the observations and improving the reporting of findings which is currently messy especially for the post-questionnaire part.

As a minor point, providing participants IDs (or the group ID) next to quotes can provide useful context to the reader (eg., to know whether the group reached agreement or not), and thus provides a richer narrative.
