Thank you for reviewing our paper. We acknowledge the valuable feedback summarized by the 1AC and our rebuttal will argue against these points: we see them as possible extensions that stem from differences in research preferences and cannot be the basis for rejecting this work. We adopt this stance because a prior round of expert reviewers at CHI identified tangible revisions, which we addressed and are confident that we can address all the improvements suggested by R1 and include the additional information requested by R2, which will not reveal any flaws (why between-subject, add t-test on likert scale of satisfaction, etc).

Focused domain. Studying collaborative search and agreement is challenging due to the complex group dynamics. By having the domain as a controlled variable, we were able to surface insights on these dynamics albeit domain-specific ones. Note the following papers that made significant contributions to collaborative search while focusing on a single search topic: [A-1] (restaurants) and ref [11](social event coordination). In Section 5, we generalize CREST to other domains like group investments, charitable investments, and event planning. We disagree that our domain is narrow and hence not worthy of focused research: as you note we motivate the problem well and will not repeat those motivations here. We also disagree that good systems research requires one to generalize design principles to multiple fields. We build on one another and if this work took two years to produce, one would expect that a rigorous process of generalizing to other domains to be equally thorough. Given the history of niche papers in UIST, we feel that this is an unfair expectation of any work.

Formative conversations. It is not unusual for formative “conversations” to be on small focused groups. We wanted to interview groups with exact and recent experience with collaboratively booking a property. The idea here is to understand how users go about it with current tools and surface needs through in-depth interviews and narratives that can then inspire our design. Validation requires larger studies (Sec 3). As R2 notes, the fact that a large body of prior work corroborates the tasks and challenges (note direct refs in every task and challenge except t5, c5 – but there are prior works that allude to these issues, which we can add) is strong evidence that we are focusing on the main ones. We can reframe the section as a distillation of related works corroborated by formative conversations.

User study. We find our *qualitative* study in line with UIST’s guidelines. The desire for more quantitative results is understandable but (a) in a multi-day async study over the web tracking time users dwell on a feature is not easy to reliably measure (e.g. clicking a search result then switching to email or making coffee) (b) time spent per feature or overall is also not a reliable measure of effectiveness in an async decision-making task, perhaps users take more time because they are seriously considering other options. An in-the-wild study with users with real bookings at stake is also ideal but an academic research lab doesn’t have access to such a platform nor can be expected to launch one. Our incentived approach is more rigorous compared to other studies (refs [26] and [A-1])that do not even attempt to ensure that in simulated scenarios users are reliably motivated to engage. Moreover, our study is entirely async unlike other simulated user studies (refs [10,11]). We do not understand why R2 views negotiation as distinct from collaboration. Negotiation was not expected nor explicity requested. Through the many features of CREST, we observe information sharing regarding preferences, which is part of collaboration. Balancing personal interest (via bonus) with the goal of successful agreement is crucial otherwise participants can quickly agree to anything to conclude the study. Finally, it is not clear how to better align our study with the research goals (R2): we qualitatively explore how users use CREST's novel features address the tasks and challenges, we explore the influence of the mediation messages, etc.

Participant diversity. WEIRD does not apply. Our campus has 110 nationalities with countries that are neither rich, industralized nor democratic. Our participants include staff, spouses, etc that are not academics in any form and alum who are no longer in the university but are reachable through a mailing list. Even so, the expats and students within this community have a strong need for finding shared housing (see Sec 1) and as such are an appropriate target group for evaluation.

Rigor. We can include t-tests on likert scale questions. We are more comfortable showing the full distribution through a spark chart (we include std-dev) as significance tests on inherently qualitative data is conventional but not rigorous. In a pilot, within-subjects test, we observed fatigue in the second run with users dropping out in the second iteration regardless of tool. Table 5 illustrates that we surpass prior works in terms of study size. We find that most reviewers want straightforward clarifications and we can add those.

We do not understand R2's comment on mismatch. There are lots of papers that focus specifically on extracting individual search preferences. These are complimentary. We clearly describe why works that focus solely on using these preference rankings to algorithmically find solutions are not ideal (Secs 2.3.c4 & 4). What we show clearly is that users can specify their preferences and through discussions and agreement features they can express how strict or soft these preferences are to reach satisfying solutions. So CREST's design captures this well.

IRB. Obtained and IRB-exempt.
Compensation. ~14USD for participation and ~27USD shopper bonus.

[A-1] https://dl.acm.org/doi/pdf/10.1145/3106426.3106505

Brainstorming

Finally, we point out that we found the AC reviews demoralizing in 2 regards: First, hinting that we should forgo a rebuttal because as it stands we have no chance makes it seem that the ACs have no intention of taking our work or rebuttal seriously. Second, suggesting that UIST is a competitive conference even though the papers address very different problems is counter to the spirit and culture of UIST. Even so, prior to a PC meeting and calibration, it is surprising to have a holistic sense of the entire program and where our work stands. We understand that it is not the intention of the ACs to demoralize and provide quality feedback which we appreciate and respect but in the absence of a face-to-face discussion and a sharing of ideas, words on paper matter.

Formative study

Too focused domain (2AC, AC)

The reviewers noted that our paper is " well motivatedmotivate and written", and our research direction is "interesting and needed". We will respond to the main concerns below.

1. Formative Study Concerns

We note that formative conversations constitute only a small part of a larger, comprehensive design process with several rounds of prototying and soliciting user feedback through pilot studies. We did not write about this but in total we had 8 iterations where we built 3 prototypes and received the feedback of 26 users. For brevity, we focused on the initial, inspiring seed conversations: it is not unusual to have formative conversations with small focus groups, but validate one's work through larger user studies. We point out that we call them "conversations" and do not frame this is as a full-fledged user "study" This process (we are happy to elaborate on our many prototypes) combined with a thorough distillationon of findings from a large body of relevant literature was how we put forward the tasks, challenges and principles. Thus, our analysis does not suffer from being informed by a skewed group and as we clearly write, and is corroborated by the prior work.

We note the following award-winning papers that motivated the design of their systems based on only on the analysis of existing literature or a motivating user in mind:
https://dl.acm.org/doi/pdf/10.1145/3526113.3545623 ,
https://dl.acm.org/doi/pdf/10.1145/3379337.3415875 ,
https://dl.acm.org/doi/pdf/10.1145/3526113.3545703,
https://dl.acm.org/doi/pdf/10.1145/3332165.3347866 .

2. Too focused domain

Collaborative search and agreement is an interesting research problem due to the complex group dynamics that emerge throughout the process. By having the domain as a controlled variable in our investigation, we were able to specifically surface insights on these dynamics albeit domain-specific ones. Note the following papers that made significant contributions to collaborative search while focusing on a single search topic:
https://dl.acm.org/doi/pdf/10.1145/3106426.3106505 (restaurants) and
https://doi.org/10.1145/3359208 (social event coordination) . Furthermore, we dedicate an entire section where we extend the lessons learned from CREST into other domains like group investments, charitable investments, and event planning. We disagree that our domain is too narrow and hence not worthy of focused research: as you note we motivate the problem well and will not repeat those motivations here. We also disagree that good systems research requires one to generalize design principles to multiple fields. We build on one another and if this work took two years to produce, one would expect that a scientific process of generalizing to other domains to be equally thorough. Given the vast history of niche areas in UIST, we feel that this is an unfair expectation of any work.

3. Implementation and Design

The expression of soft vs strict preferences by the user is one that we decide not to focus on in our system. In particular, the Team's Preferences panel allows users to visually express their preferences and budgetary requirements and the user icon of each user appears next to the preferences they like. We note that the collaborative search process is an ever-evolving one, especially in the multi-day scenario we try to capture. As such, a search session in CREST begins with the assumption that every preference is important and is bound to change over time as users begin communicating with each other and they receive recommendations from CREST-bot, our automated mediation agent. Assigning a numerical value to a preference, or a discrete classification, lends itself more readily to algorithmic solutions to collaborative search, solutions that might varying satisfaction levels depending on the algorithm employed (https://dl.acm.org/doi/10.1145/3131361). Although such solutions are interesting to consider, that is not the focus of this work. This design decision can be further explored in our revised manuscript.

4. User study

We note that IRB approval was obtained for this study and that participants were compensated 14 USD for a 2 hour total commitment, with a 27 USD bonus available to one of the participants in the group that best represents the requirements of their avatar. A valid concern with studies within college campuses is that participants are WEIRD. Our participants, however, were not only students. Our global campus is not Western (110 nationalities) and many of ours students, employees, etc. come from poor, non-industralized, non-democratic countries. They are educated and comfortable with technology. These demographics aside, this is a case where the college campus provides an appropriate pool of users for the system we are designing: students and researchers do frequently look for shared accommodations and many within our community are expats and hence do look for places to live with family together over breaks, for example or need to find cheap long-term accomodations in an expensive city. In terms of rigor, we compare our user study with others on collaborative search in Table 5 and find that ours surpasses all but 2 user studies in size. Ours is the only one that acknowledges head on the limits of replicating group dynamics in simulated environments and addresses it through a personal-shopper expirement. Furthermore, note that following UIST paper that carried out their user studies in college campuses: https://dl.acm.org/doi/pdf/10.1145/3526113.3545619.

Although our discussion section focuses on deriving insights from the interesting group dynamics that emerged in our user studies, our contributions are entirely UI/UX focused with a particular focus on applicability and usability. As such, we measured how usable our system is through a likert-scale post-experiment questionnaire, where each likert-scale question was followed by a a question asking users to justify their rating. In this questionnaire, users provided positive feedback on the design of our UI, specifically praising how specific UI elements like the collabo-ratio visualization are " …great at illustrating who has more say in the decision-making which is very interesting", to name a few. Again, an approach utilized by UIST best paper awardees (https://dl.acm.org/doi/pdf/10.1145/3526113.3545620). Furthermore, evaluating the usage of a system through a usability study and collecting feedback from the users through a likert style questionnaire and open-ended questions are methods promoted by the UIST chairs in their author guide (see https://dl.acm.org/doi/pdf/10.1145/3173574.3173610). If needed, we can include the full post-experiment questionnaire, a more in-depth statistical analysis of the likert-scale responses (t-tests included), and more comment excerpts in the appendix of the revised manuscript.

Interesting papers we might want to mention:
https://dl.acm.org/doi/pdf/10.1145/3526113.3545623 ,
https://dl.acm.org/doi/pdf/10.1145/3526113.3545703 ,
https://dl.acm.org/doi/pdf/10.1145/3472749.3474800

CREST UIST Rebuttal

Thank you for reviewing our paper and your valuable feedback. The reviewers noted that our paper is " well motivate and written", and our research direction is "interesting and needed". We will respond to the main concerns below.

5. Formative Study Concerns

We note that formative conversations constitute only a small part of a larger, comprehensive design process with several rounds of prototyping and soliciting user feedback through pilot studies. We did not write about this but in total we had 8 iterations where we built 3 prototypes and received the feedback of 26 users. For brevity, we focused on the initial, inspiring seed conversations: it is not unusual to have formative conversations with small focus groups, but validate one's work through larger user studies. We point out that we call them "conversations" and do not frame this is as a full-fledged user "study" This process (we are happy to elaborate on our many prototypes) combined with a thorough distillationon of findings from a large body of relevant literature was how we put forward the tasks, challenges and principles. Thus, our analysis does not suffer from being informed by a skewed group and as we clearly write, and is corroborated by the prior work.

We note the following award-winning papers that motivated the design of their systems based on only on the analysis of existing literature or a motivating user in mind:
https://dl.acm.org/doi/pdf/10.1145/3526113.3545623 ,
https://dl.acm.org/doi/pdf/10.1145/3379337.3415875 ,
https://dl.acm.org/doi/pdf/10.1145/3526113.3545703,
https://dl.acm.org/doi/pdf/10.1145/3332165.3347866 .

6. Too focused domain

Collaborative search and agreement is an interesting research problem due to the complex group dynamics that emerge throughout the process. By having the domain as a controlled variable in our investigation, we were able to specifically surface insights on these dynamics albeit domain-specific ones. Note the following papers that made significant contributions to collaborative search while focusing on a single search topic:
https://dl.acm.org/doi/pdf/10.1145/3106426.3106505 (restaurants) and
https://doi.org/10.1145/3359208 (social event coordination) . Furthermore, we dedicate an entire section where we extend the lessons learned from CREST into other domains like group investments, charitable investments, and event planning. We disagree that our domain is too narrow and hence not worthy of focused research: as you note we motivate the problem well and will not repeat those motivations here. We also disagree that good systems research requires one to generalize design principles to multiple fields. We build on one another and if this work took two years to produce, one would expect that a scientific process of generalizing to other domains to be equally thorough. Given the vast history of niche areas in UIST, we feel that this is an unfair expectation of any work.

7. Implementation and Design

The expression of soft vs strict preferences by the user is one that we decide not to focus on in our system. In particular, the Team's Preferences panel allows users to visually express their preferences and budgetary requirements and the user icon of each user appears next to the preferences they like. We note that the collaborative search process is an ever-evolving one, especially in the multi-day scenario we try to capture. As such, a search session in CREST begins with the assumption that every preference is important and is bound to change over time as users begin communicating with each other and they receive recommendations from CREST-bot, our automated mediation agent. Assigning a numerical value to a preference, or a discrete classification, lends itself more readily to algorithmic solutions to collaborative search, solutions that might varying satisfaction levels depending on the algorithm employed (https://dl.acm.org/doi/10.1145/3131361). Although such solutions are interesting to consider, that is not the focus of this work. This design decision can be further explored in our revised manuscript.

8. User study

We note that IRB approval was obtained for this study and that participants were compensated 14 USD for a 2 hour total commitment, with a 27 USD bonus available to one of the participants in the group that best represents the requirements of their avatar. A valid concern with studies within college campuses is that participants are WEIRD. Our participants, however, were not only students. Our global campus is not Western (110 nationalities) and many of ours students, employees, etc. come from poor, non-industralized, non-democratic countries. They are educated and comfortable with technology. These demographics aside, this is a case where the college campus provides an appropriate pool of users for the system we are designing: students and researchers do frequently look for shared accommodations and many within our community are expats and hence do look for places to live with family together over breaks, for example or need to find cheap long-term accomodations in an expensive city. In terms of rigor, we compare our user study with others on collaborative search in Table 5 and find that ours surpasses all but 2 user studies in size. Ours is the only one that acknowledges head on the limits of replicating group dynamics in simulated environments and addresses it through a personal-shopper expirement. Furthermore, note that following UIST paper that carried out their user studies in college campuses: https://dl.acm.org/doi/pdf/10.1145/3526113.3545619.

Although our discussion section focuses on deriving insights from the interesting group dynamics that emerged in our user studies, our contributions are entirely UI/UX focused with a particular focus on applicability and usability. As such, we measured how usable our system is through a likert-scale post-experiment questionnaire, where each likert-scale question was followed by a a question asking users to justify their rating. In this questionnaire, users provided positive feedback on the design of our UI, specifically praising how specific UI elements like the collabo-ratio visualization are " …great at illustrating who has more say in the decision-making which is very interesting", to name a few. Again, an approach utilized by UIST best paper awardees (https://dl.acm.org/doi/pdf/10.1145/3526113.3545620). Furthermore, evaluating the usage of a system through a usability study and collecting feedback from the users through a likert style questionnaire and open-ended questions are methods promoted by the UIST chairs in their author guide (see https://dl.acm.org/doi/pdf/10.1145/3173574.3173610). If needed, we can include the full post-experiment questionnaire, a more in-depth statistical analysis of the likert-scale responses (t-tests included), and more comment excerpts in the appendix of the revised manuscript.

Interesting papers we might want to mention:
https://dl.acm.org/doi/pdf/10.1145/3526113.3545623 ,
https://dl.acm.org/doi/pdf/10.1145/3526113.3545703 ,
https://dl.acm.org/doi/pdf/10.1145/3472749.3474800