Point-by-point Responses to the Reviewers' Comments (Rd. 2)

Again, we would like to express our sincere gratitude for the time and effort to review our manuscript.

Reviewer 1

1. An error in the power analysis R code

I think the authors made a small mistake in their code to determine the sample size for their multilevel models. Lines 32 and 33 of their script power_analysis_H12.R read:

> *r.y <- rnorm(Npart*Nprof, sqrt(1-ICC))*
>
> *r.med <- rnorm(Npart*Nprof, sqrt(1-ICC))*

This is equivalent to

> *r.y <- rnorm(Npart*Nprof, mean = sqrt(1-ICC), sd = 1)*
>
> *r.med <- rnorm(Npart*Nprof, mean = sqrt(1-ICC), sd = 1)*

which is probably not what the authors intended. (Note that the same error is also present on lines 27 and 28 in analysis_code.R.)

Instead the code should probably read:

> *r.y <- rnorm(Npart*Nprof, 0, sqrt(1-ICC))*
> *r.med <- rnorm(Npart*Nprof, 0, sqrt(1-ICC))*

Thank you for pointing out this critical error. We corrected "*rnorm(Npart*Nprof, sqrt(1-ICC))*" to "*rnorm(Npart*Nprof, 0, sqrt(1-ICC))*" and added session information. Please refer to the new code on the OSF (html: https://osf.io/uep8f, R: https://osf.io/82f9t).

2. Removing profiles that fail the manipulation check in the main experiment

The authors report the following with respect to the analyses of the main experiment (page 16):

Pairs of profiles whose perceived abstractness do not differ significantly by condition will be

excluded from the analysis.

I do not think that this choice is a good idea, but I am not completely sure about that: The

authors already use their preliminary survey (which uses a different sample of participants) to select appropriate profiles. If (in the main experiment) they again exclude profiles that fail the manipulation check (in the main experiment), I would worry that this limits the generalizability of their findings. The only plausible reasons apart from a type II error, why the profiles "work" in the preliminary survey but do not work again in the main experiment, would have to be attributed to sample characteristics of the participants in the main experiment. I totally agree that having the manipulation check present in the main experiment is a good idea, and if they want, the authors can also add the analysis without the profiles that fail the manipulation check (in the main experiment) as an additional sensitivity analysis. However, I think that stronger claims (in the sense of a more severe test) can be made from the results of the main experiment if the preregistered analyses for H1 and H2 include all profiles that were chosen to be appropriate in the preliminary survey.

Thank you for your comment. We had included the manipulation check in the main experiment because stimuli of the main experiment (profile) are different from the preliminary study (sentence), as you suggested in Rd. 1. However, we also agree with your concern that excluding profiles that fail manipulation checks will limit generalizability. Thus, we decided to use all profiles in the pre-registered analyses and add an exploratory analysis without profiles that fail manipulation checks.

"*First, we will check whether our manipulation worked successfully by conducting one-tailed t-tests of the perceived abstractness of ten pairs of profiles. Although all pairs will be included in the hypothesis-testing analyses regardless whether the differences of perceived abstractness of profile turn out significant, we will conduct additional sensitivity analysis that exclude pairs of profiles whose perceived abstractness do not differ significantly by condition.*"

(lines 359 - 364)

3. Decide whether testing H2 dependent on the result for H1

The authors report the following with respect to the analyses in the main experiment (page 16):

Only if H1 is supported, we will proceed to test H2 by conducting a multilevel mediation analysis on attraction with abstractness, random effects of participants as predictors, and attributional confidence as a mediator.

I do not think that it is a good idea to test mediation only if a total effect of abstractness on attraction is confirmed. Although the older literature on mediation claims the contrary, it is theoretically possible for the indirect effect to perfectly cancel out the direct effect, which could result in a pattern where the total effect is 0 but the indirect effect (and the direct effect) is unequal to 0. Additionally, because any decision against H1 can always be a type II error, I would strongly suggest to always report the result of the mediation analysis (H2), irrespective of the result for the total effect (H1).

Thank you for your suggestion. We changed the previous plan to test H2 (mediating effect) even if H1 (direct effect) was not supported.

"*Regardless of whether H1 is supported, we will proceed to test H2 by conducting a multilevel mediation analysis on attraction with abstractness, random intercepts of participants as predictors, and attributional confidence as a mediator.*"

(lines 371 - 373)

(1) Justification of effect sizes and variances for power analysis

Although I think the most important task is to make the assumptions for power calculations transparent (which is the case here, because the manscript together with the provided R code includes all assumptions), the authors could give some short justification on how these values have been chosen. This includes the following parameters:

- The effect size, margins, standard deviation, and equivalence bounds for the equivalence tests of the preliminary survey.

- The ICC values (mediator and outcome model), and the effect sizes for the three path coefficients in the mediation model.

Thank you for your helpful comment. We added some justifications below to the manuscript.

<u>Preliminary survey</u>

- **<u>Effect size, Equivalence bounds (Δ)</u>**
  We set our minimally interesting effect size as Cohen's *d* = 0.3 throughout the preliminary survey. Thus, we set the target effect size in t-tests as *d* = 0.3, and equivalence bounds in TOST as Δ = 0.33 (corresponds to *SD* = 1.1).
- **<u>SD</u>**
  We conducted a tiny pilot study to estimate the SD value. In the pilot study, ten (graduate) students rated favorability of four sentences extracted from the stimuli of the preliminary study. The mean of SD for four sentences was 1.1.

<u>Main study</u>

- **<u>Effect size</u>**
  We assume effect sizes in the main study as β = .51 (uncertainty to attraction) and β = .34 (manipulation to uncertainty), respectively. The former is based on a past study (Baruh & Cemalcılar, 2018). Regarding the latter, due to the lack of study that adopts the same manipulation as ours, we could not assume a specific value. Instead, we speculated the effect size will be between small and medium, considering that we confirmed that there are certain differences (*d* < 0.3) in the abstractness of stimuli of both conditions in the preliminary study.
- **<u>ICC</u>**
  We could not estimate a specific ICC value because no previous study adopted similar stimuli to this study. However, we speculated that it will be at least a moderate size because the ratings for all profiles will be made by the same individuals. Thus, we assumed ICC as 0.6 in our simulation. Although we could assume higher values (e.g., 0.8), it is equivalent to assuming that our manipulation has little or no effect, so we thought it was meaningless.

"*We determined the target effect sizes based on our minimally interested effect size in the preliminary study: Cohen's d = 0.3. The assumed value for SD of favorability ratings, 1.1, was derived from our small pilot study.*"

(lines 228 - 229)

"*The assumed effect size of uncertainty on attraction (β = .51) was determined based on a previous study (Baruh & Cemalcılar, 2018). For the effect of experimental manipulation on uncertainty, we could not assume a specific value due to the lack of study that adopts the same manipulation as ours. Instead, we speculated the effect size will be between small and medium (β = .34), considering that we confirmed that there are certain differences in the abstractness of stimuli (d < 0.3) in the preliminary survey. Similarly, we could not estimate a specific ICC value because no previous study adopted similar stimuli to this study. However, we speculated that it will be at least a moderate size because the ratings for all profiles will be made by the same individuals. Thus, we assumed ICC as 0.6 in our simulation. Although we could assume higher values (e.g., 0.8), it is equivalent to assuming that our manipulation has little or no effect, thus we thought it was meaningless.*"

(lines 299 - 309)

(2) Position of the manipulation check items in the main experiment

On page 14 of the manuscript, the authors write with respect to the main experiment: *Participants will answer one manipulation check item (“How abstract did you feel the profiles were?”; 7 points scale: “abstract” - “concrete”) for each profile.*

Based on this description, it is not clear to me, whether this item is presented directly after each profile or at the end of the questionnaire together for all profiles. Both options probably have advantages and disadvantages. Personally, I think the stronger design would be to place all manipulation checks at the end of the questionnaire, to make sure that being asked explicitly about the abstractness of a profile does not change the response to the following profiles (because the abstractness of the profiles has been made more salient by the manipulation check of previous profiles)

Thank you for your comment. We initially planned to place the manipulation check item directly after each profile. However, your concern that this procedure may provoke demand characteristics is plausible. We decided to change the previous plan to place all manipulation check items at the end of the questionnaire.

“*All manipulation check items will be placed at the end of the questionnaire to avoid demand characteristics.*”

(lines 351 - 353)

(3) Document software versions

- As already mentioned earlier, I would encourage the authors to document the software versions they used when running their power analysis.
- The authors write in their manuscript that they will use R version 4.3.3 for their final analysis. Although preregistering the R version is of course fine, I think it is not absolutely necessary and I would not see it as a problem if the authors use a later R version for their final analysis, as long as they document this R version in their stage 2 manuscript.

Thank you for your suggestion. We reported software versions used in the power analysis (https://osf.io/uep8f).

(4) Preregister R code for equivalence tests

While the authors have provided R code for their final mediation analysis, they have not provided code on how to perform the equivalence tests in their preliminary survey, and how their power analysis for these equivalence tests has been conducted. Although not absolutely necessary, the authors could (for transparency reasons) also upload the code for running the planned equivalence tests, to ensure that there are no unspecified analysis settings that could be considered as unnecessary researcher degrees of freedom.

Thank you for your suggestion. We uploaded the analysis codes for the preliminary survey to the OSF (power analysis: https://osf.io/qsdrv, analysis: https://osf.io/ybwmn).

(5) Mention extended analysis options in a limitation section

Although I consider the preregistered analyses appropriate, I have thought about possible limitations of the current statistical approach and I want to briefly report my thoughts here:

- The authors use multilevel models with a random intercept for participants but no random intercept for profiles. I think this choice is acceptable because the mediation R package cannot handle multilevel model with multiple random intercepts. However, without this constraint I would consider a model with random intercepts for both participants and profiles even more appropriate. Although partly speculative, I would suspect that such an extended model might be able to control for unobserved confounding between the mediator and the outcome that is caused by attributes of the individual profiles. For this reason, it might be reasonable to include the lack of modeling random intercepts for profiles in the limitation section of the stage 2 manuscript. (Side comment: Running mediation analyses with more complicated models would theoretically be possible with the brms R package. But this would require computing the mediation effect "manually", so I think this is not worth the effort for the current study.)

- The authors do not plan to run sensitivity analyses against potential unobserved confounding between the mediator and the outcome. I think this choice is acceptable because the medsens() function in the mediation cannot handle multilevel models. I think it might be reasonable to include the lack of sensitivity analyses in the limitation section of the stage 2 manuscript. (Side comment: Running sensitivity analyses with multilevel mediations models would theoretically be possible with the brms R package. But this would require computing the mediation effect "manually", so I think this is not worth the effort for the current study.)

Thank you for your valuable comments. We agree with your thought that (1) not considering random effects of profiles and (2) not running sensitivity analysis are limitations of our study. We will write the stage 2 manuscript taking into account those points.

(6) Suggestions to further improve their R code

I have some minor comments on the R code in power_analysis_H12.R that do not affect the performance of the script:

- In line 66 and 67, the authors define the variables VARR<-1 and ALPHA<-0.05 that are not used anywhere else in the script.

- In line 7, the authors run rm(list=ls()). Although this is not problematic by itself, I just wanted to make the authors aware that this practice is considered error-prone by many R programmers, because it does not ensure a clean R session and can therefor affect reproducibility of results.

Thank you for pointing out problems with the R code. We deleted "VARR <- 1", "ALPHA <- 0.05", and "rm(list=ls())" from the simulation code.

(7) Iterations for the quasi-Bayesian confidence interval

The authors report that they will run 10000 resamples for the quasi-Bayesian confidence interval of their final mediation analysis. Although more iterations are of course always better, this number sounds a bit excessive. It is so high that I could not run the analysis on my laptop with 16 GB memory. Perhaps 5000 or 2000 iterations might also be enough, but I leave this decision to the authors.

Thank you for your suggestion. We corrected the number of resamples from 10000 to 5000.

"*To estimate the indirect effect, we will employ a quasi-Bayesian confidence interval (5,000 resamples).*"

(lines 373 - 374)

(8) Improve wording and fix small mistakes

- On page 16 of the manuscript, the authors write:
  To test H1, we will include abstractness (dummy variable: abstract condition = 0, concrete condition = 1) and random effects of participants as predictors, with attraction as the dependent variable. [...] Only if H1 is supported, we will proceed to test H2 by conducting a multilevel mediation analysis on attraction with abstractness, random effects of 368 participants as predictors, and attributional confidence as a mediator.
  I would suggest to use the more precise term "random intercepts" instead of "random effects" here.

- On page 14 of the manuscript, the authors write:
  One item of DQS (Directed Questions Scale; Maniaci & Rogge, 2014) for each profile such as "Choose 1 in this question." were operated in order to detect participation with paying attention (i.e., satisficers).
  I think what the authors want to say is that these items are used to detect participants that do NOT pay attention.

- On page 11 of the manuscript, the authors write:
  During stimulus selection, we will carefully consider their semantic proximity, as it may impact the perceived consistency of the target person.
  I am not completely sure what the authors mean by this sentence, perhaps they could add another sentence to make clear how they plan to control for semantic proximity when selecting the stimuli.

Thank you for your suggestion. For the first point, we replaced the word "*random effects*" of "*random intercepts*". For the second point, we corrected "*participation with paying attention*" to "*participation without paying attention*". For the last point, we added some sentences as follows.

"*We will create ten profiles using 20 pairs of sentences that are selected through the above analysis. When creating profiles, we will carefully select sentences not to diminish the consistency of personality of the depicted person and not to include too private information.*"

(lines 280 - 282)

Reviewer 2

1)

As the preliminary study is pre-registered it needs its entries in a design table.

Thank you for your comment. We added a design table for the preliminary survey.

2) p 10

We calculated the required sample size for a paired t-test under the assumption of α= .05 and Cohen's d = 0.2, and the Two One-Sided Test (TOST; Schuirmann, 1987) using the PowerTOST package (Labes et al., 2024) in R. Under the assumption of α = .05, the margin(Δ) = 0.3, and a standard deviation of 1.0, our analysis showewd that 156 and 215 participants,respectively, would suffice to achieve 80% statistical power.

It is not clear where these numbers come from. Why d = 0.2? Why SD = 1? Why delta = 0.3? The authors do not make explicit why minimally interesting effect sizes are 0.2 in one case, 0.3 in another. Further, to be inferentially consistent, as they use "inference by intervals" for favourability ratings, it would be consistent to do so for abstractness as well, that is find matched stimuli whose abstractness differed by more than a minimally interesting amount.

The meaning of a d of 0.2 or 0.3 is hard to intuit for these sentences. I think it is worth running a small pilot, just to estimate very roughly the SD of the ratings, so that the minimally interesting effects can be set in raw units, and power calculated for that null region.

Thank you for your thoughtful comment. As you pointed out, the rationale of the assumption about parameters was insufficient. We added some discussions on three issues you have identified, as follows.

1. **Minimally interesting effects for t-test and that for TOST are unequal**
   Thank you for your comment. We standardized the minimally interesting effect size as *d* = 0.3 throughout the preliminary study.

2. **The authors should use "inference by intervals" like TOST for favorability ratings**
   Thank you for your comment. We consider that using both t-test and TOST (two one-sided t-tests) is not problematic. This is because TOST is not "inference by intervals" as it simply repeats two t-tests and thus uses *p* value for testing hypotheses.

3. **It is worth conducting a small pilot to roughly estimate the SD value**
   Thank you for your suggestion. We conducted a tiny pilot study to estimate the SD value. In the pilot study, ten (graduate) students rated favorability of four sentences extracted from the stimuli of the preliminary study. The mean of SD for four sentences was 1.1. Thus, we re-performed the power analysis assuming SD = 1.1, upper and lower bounds (delta) = 0.33 (corresponds to *d* = 0.3).

"A total of 250 adult native Japanese speakers recruited from a crowdsourcing platform will participate in the survey in exchange for monetary compensation. We calculated the required sample size for a paired t-test under the assumption of α = .05 and Cohen's d = 0.3, and for the Two One-Sided Test (TOST; Schuirmann, 1987) under the assumption of α = .05, the upper and lower bounds (Δ) = 0.33, and SD of 1.1. We used the PowerTOST package (Labes et al., 2024) in R to perform the power analysis for TOST. The analyses showed that 71 and 192 participants, respectively, would suffice to achieve 80% statistical power. Consequently, we decided to collect

***data from 250 participants in the preliminary study. We determined the target effect size based on our minimally interested effect size in the preliminary study: Cohen's d = 0.3. The assumed value for SD of favorability ratings, 1.1, was derived from our small pilot study.*"**

(lines 220 - 230)

3) Main study:

There is no justification for what the minimally interesting effect is that they do not want to miss out on detecting. A scientific reason needs to be given for what effect is just worth missing out on, and that value used in power calculations. (I find *predicted* effects easier to scientifically justify and hence use Bayes factors, e.g. for mediation and testing differences: https://doi.org/10.1177/2515245919876960. But it is of course up to the authors which inferential route they take.)

Thank you for your comment. In the original power analysis, we used the medium effect size (r = .39) for paths both from uncertainty to attraction and from experimental manipulation to uncertainty. However, we reconsidered those values and changed them to β = .51 (uncertainty to attraction) and β = .34 (manipulation to uncertainty). The former is based on a past study (Baruh & Cemalcılar, 2018). Regarding the latter, due to the lack of study that adopts the same manipulation as ours, we could not assume a specific value. Instead, we speculated the effect size will be between small and medium, considering that we confirmed that there are certain differences (*d* < 0.3) in the abstractness of stimuli of both conditions in the preliminary study. Thus, we determined the effect size as β = .34. We added those discussions to the manuscript, as follows.

"*The assumed effect size of uncertainty on attraction (β = .51) was determined based on a previous study (Baruh & Cemalcılar, 2018). For the effect of experimental manipulation on uncertainty, we could not assume a specific value due to the lack of study that adopts the same manipulation as ours. Instead, we speculated the effect size will be between small and medium (β = .34), considering that we confirmed that there are certain differences in the abstractness of stimuli (d < 0.3) in the preliminary survey. Similarly, we could not estimate a specific ICC value because no previous study adopted similar stimuli to this study. However, we speculated that it will be at least a moderate size because the ratings for all profiles will be made by the same individuals. Thus, we assumed ICC as 0.6 in our simulation. Although we could assume higher values (e.g., 0.8), it is equivalent to assuming that our manipulation has little or no effect, so we thought it was meaningless.*"

(lines 299 - 309)

4) Final column of design table:

State in here with a simple proposition the most general claim that the test could find evidence against. Maybe it is something a bit more specific than uncertainty reduction theory. But why isn't uncertainty reduction theory challenged by finding no difference in attraction between abstract and concrete profiles? Why would that theory not predict a difference in this study? Be explicit about this.

1. **<u>State in here with a simple proposition the most general claim that the test could find evidence against</u>**
   Thank you for your comment. We stated a simple proposition that our research could show is wrong, as follows.

“*If our hypotheses are not supported, the results will show the following proposition are wrong: “More concrete expressions in online profiles contribute to uncertainty reduction, resulting in increased attraction of the target person.”*”
(p 19)

2. **Why isn't uncertainty reduction theory challenged by finding no difference in attraction between abstract and concrete profiles?**
Thank you for your comment. If the result showed that uncertainty did not correlate with attraction, it indeed seems to contradict the theory. However, the situation dealt with in this study, in which participants one-sidedly rate the target person without interactions, differs from the situation that URT originally assumed. Therefore, even if we obtain a contrary result, it will not necessarily mean the URT itself is invalid. We added those discussion, as follows.

“*However, even if we obtain such a contrary result, it will not necessarily mean the URT is invalid. This is because the situation dealt with in this study, in which participants one-sidedly rate the target person without interactions, differs from the situation that URT originally assumed.*”
(p 19 -20)

5) I still find the following alternative theory a highly plausible competing theory that makes the same predictions for this study:

The more certain one is of someone's positive nature, the more one is attracted to them. The authors make clear that this theory is a different theory to uncertainty reduction theory, the latter being what they are interested in. It is good to be clear about that, but that means obtaining evidence in support of uncertainty reduction theory won't actually provide much support for specifically that theory. However, falsifying the rpediction would count against uncertainty reduction theory, as far as these two theories are concerned. The latter point still makes the study worthwhile. Maybe the thing to do is just be clear about this. Or else use only items that are very slightly disfavourable, as the theories can make contrasting predictions then.

Thank you for your comment. Although we only use positive sentences as experimental stimuli, our interest is the effects of uncertainty itself, not the sub-theory of URT that applies only to positive information. We did not include negative sentences to the stimuli because concretely indicating norm violation of the target person will diminish the person’s attraction and thus hide the effect of uncertainty reduction. Although it is a necessary step to correctly measure the effect of abstractness, we consider that not using negative sentences is one of the limitations of our study, and we will discuss this point in the Stage 2 manuscript.

6) It also occurs to me there is another theory that is relevant, namely relevance theory (https://en.wikipedia.org/wiki/Relevance_theory) or, more generally, pragmatics (e.g.https://plato.stanford.edu/entries/pragmatics/) .

Consider the Gricean axioms (quoted from last link):

"Make your contribution as informative as is required (for the current purposes of the exchange).

Do not make your contribution more informative than is required."

In the context of these profiles, there may be presumptions of the sort of detail that is typically used or appropriate. If someone deviates from that level of detail, the pragmatics of the communication will feel strange. The authors give statements in the abstract that may appear too abstract given they easily could have been more concrete for the same length. Conversely the axioms predict a profile could present "more than I need to know". It would be an empirical matter where the sweet spot was, and how far from it and in which direction the authors' examples were. (Could be tested by rating the concreteness/abstractness of real profiles on the same scale as the authors use and seeing whether the authors' profiles are more or less concrete than the prototype.)

This theory strikes me as plausible. If typical relevance concerns are pragmatically violated more in the abstractness than concrete condition, the profile may come across as odd, and hence untrustworthy. Or it may be vice versa; in which case an outcome falsifying uncertainty reduction theory may not actually count against it. Something should be done about this concern - minimally, discussing it - but it may be worth doing more than that.

Thank you for your valuable comment. We agree with your thought that too much information can negatively affect attraction. Past literature indeed suggested that too much disclosure of personal information diminishes attraction, if the disclosure is inconsistent with social norms (Altman & Taylor, 1973; Collins & Miller, 1994). However, we consider that this problem will not confound our result because the amount of information we provide participants in our experiment is small (i.e., only 5 short sentences, which is fewer than a past study (Baruh & Cemalcılar, 2018)), and because we will create the stimuli not to include too private information, not to violate social norms. Thus, based on URT, we can certainly predict an increase of the information amount in our experimental setting will lead to more attraction. We discuss this problem in the revised manuscript as follows, and will add some discussions in the Stage 2 manuscript.

"*It is plausible that excessive certainty can dampen interest and fail to engender attraction in more mature relationships. Empirical evidence indeed demonstrated that ambiguous affection is more likely to heighten romantic attraction than unambiguous affection (Whitchurch et al., 2011) and that uncertainty increased arousal (Greco & Roger, 2003; Ramsøy et al., 2012), which contributes to heightened attraction (Foster et al., 1998; Lewandowski & Aron, 2004; Little et al., 2014). However, in the early stages of a relationship—when uncertainty prevails—it is reasonable to posit that uncertainty reduction will contribute to heightened attraction before people get bored.*"

(lines 114 - 121)