Manuscript No. XGE-2019-1738R1  
The Shared Features Principle: If Two Objects Share a Feature, People Assume Those Objects Also Share Other Features  
Journal of Experimental Psychology: General  
   
Dear Dr. Hughes,  
   
I have received three reviews of the manuscript entitled “The Shared Features Principle: If Two Objects Share a Feature, People Assume Those Objects Also Share Other Features” (XGE-2019-1738R1) that you recently submitted to Journal of Experimental Psychology: General. I sent this revision to the same reviewers who read the first version of your paper. You will find their reviews below. I read the manuscript and your response letter prior to receiving these reviews in order to gain an independent perspective on the paper, and then again with the reviews in hand. As you will see, the Reviewers recognize your responsiveness (see Reviewer 1’s comments), and I was myself impressed by the extensive additions you made to the previous version with three new experiments. However, although your revision constitutes a significant improvement over the previous version, the Reviewers still raise some points that should be addressed in order to improve an already excellent paper.

**Authors**: We thank the Editor as well as the Reviewers for their continued feedback and input on the paper. We have further revised the paper in light of those suggestions (*see below*).

**Editor**: Thus, I would like to encourage you to submit a further revision. Among the points raised by the reviewers, and which must all be carefully taken into account, the following ones particularly caught my attention.

Reviewer 2 questions the success of the EPT in Experiment 7 and notes the low effect size, suggesting that the report should not hide the weakness of the results, and I agree with the Reviewer. It is true that the replicability of this result is not warranted, but you ran so many experiments that I won’t ask you to run a replication (nor is it the wish of the Reviewer). However, as Reviewer 2 suggests, I think that the Results section should report the sequential Bayesian analysis, and just report the t values as additional information. The choice of an upper criterion of 3 for BF10 might seem unusual (it is not for me, but Reviewer 2 thinks differently and there are some debates about this point in the community) and should be motivated, as well as the criterion of .16 for BF01 (I would have expected .33 based on the criterion of 3 for BF10).

**Authors**: In line with Reviewer 2’s request, we have revised the Results section of Experiment 7. We are more transparent about the weakness of the results, report the sequential Bayesian analyses, and report the t values as additional information. We also justify our choice of upper and lower criterion for the Bayes factors as well (see revised Results section on pp. 41-43).

**Editor**: As a non-specialist of the domain, I was surprised to see the results of the EPT expressed in terms of “preference” of T1 over T2 and not in terms of priming effects. I expected to see results reporting the mean RTs for each of the four experimental conditions in order to have a clear idea of what happens in this experiment (instead of calculating differences, as it is explained on p.36, why not comparing “incongruent” trials in which prime and target have different valence to “congruent” trials in which valence is the same?).

**Authors**: In line with the Editor and Reviewer 2’s suggestions, we revised how we calculated the Evaluative Priming effect. Specifically, we now compare congruent to incongruent trials as requested (see changes on p.42).

**Editor**: Reviewer 3 identified some remaining inconsistencies in writing methods and results sections that should be fixed. I must agree that despite several readings of your work, I was myself often puzzled about the exact design of your experiments and what T1 and T2 exactly stand for. Moreover, the wording of the exploratory questions should be added in the main text as well as the fact that the responses to these questions will be treated in the meta-analysis and not before. For non-specialists of the domain, you might want to clearly explain what you understand by hypothesis and influence awareness (contingency awareness is clearer, but I’m not sure if this type of awareness is damageable for the conclusions or not – see Reviewer 3 on this latter point).

**Authors**: In line with the Editor and Reviewer 3’s suggestions, we have revised the paper further to increase clarity of our message (see revised paper and our response to Reviewer 3’s specific comments).

**Editor**: As suggested by Reviewer 1, you might try to motivate the additional experiments by methodological and theoretical considerations instead of Reviewers’ requests (it is still possible to thank these Reviewers in some footnote for the suggestion).

**Authors**: In line with Reviewer 1’s suggestions we have revised the motivation of the experiments from a reviewer suggestion to personal motivation (see changes on pp.34, 39, 43-44).

**Editor**: Of course, all the other suggestions made by the Reviewers should be taken into account.

**Authors**: We have replied to all other comments by the reviewers (*see below*).

To submit a revision, go to https://www.editorialmanager.com/xge/ and log in as an Author. You will see a menu item called "Submissions Needing Revision". You will find your submission record there. If the opportunity to revise is for some reason closed on the web when you are ready with a revision, please contact us to re-open it rather than submitting the paper as new. Also, at the top of the manuscript and cover letter, please write “Revision of XGE-2019-1738R1as invited by the action editor, Pierre Barrouillet.”  
    
Sincerely,

Pierre Barrouillet  
Associate Editor  
Journal of Experimental Psychology: General

Reviewers' comments:  
   
Reviewer #1: I am impressed with the authors' thoughtful and extensive revisions. The experiments more than sufficiently address my concerns about the odd-one-out explanation and expand the scope of the work into person perception, which were my two primary concerns. I believe this contribution to be timely given the growing interest in learning and its underlying mental representations, and this manuscript can potentially broaden our understanding and forge connections with adjacent research interests. I support publication of this article and commend the authors.  
  
I do, however, think the paper would benefit from reframing the additional experiments to be motivated by more than simply reviewer requests. While I of course understand that reviewers and authors disagree and oftentimes more work is done than what the authors believe necessary, phrases such as "based on a reviewer request" (p. 47) and "yet a reviewer suggested" (p. 30) give of the unnecessary and unwanted impression that the authors didn't really want to carry out the subsequent work when their commendable actions suggest otherwise.

**Authors**: We thank Reviewer 1 for his/her kind words and useful feedback throughout the review process. We wanted to give credit to the reviewer but understand that the phrases we used could also be understood differently. In line with the suggestions of Reviewer 1 we have now reframed the motivation for conducting the additional experiments, from a reviewer request to more personal motivation (see changes on pp.34, 39, 43-44).

Reviewer #2: I was Reviewer 2 in the first round. I accept the responses made by the authors to my criticism. However, from my point of view there remains one major issue.  
I was the Reviewer who asked for replication of the IAT results by a further implicit measure. The authors decided to replicate Exp. 4 by replacing the IAT with the evaluative priming measure (EPT). That is laudable.

On first sight, they were successful by finding a significant priming effect that corresponded to their hypotheses. But were they really successful?  
The following lines are not meant to conclude "no, not successful". However, given the low effect size in the EPT experiment, success is a least questionable. Of course, finally evaluation depends on how valid a reader finds the IAT results, that is, how much need for an alternative measure is seen by this reader. For readers like me (i.e., one who sees the IAT critical especially in the present case; see my former review), the EPT results are disappointing.

**Authors**: We appreciate the concerns of Reviewer 2 but would like to note that EPT effects in learning studies generally tend to be small and unreliable (also see our response to the next point of Reviewer 2). Moreover, we believe it is also important to consider the consistency of our results. Across seven of our eights experiments, across self-reported ratings, behavioral intentions, and two indirect procedures, across different stimuli, procedures, shared features, and samples we repeatedly find evidence for shared features effects.

**Reviewer 2**: (1) The authors report the main result as:  
"EP scores differed depending on the valence of the SO that shared a size with a TO, t(484.67) = -2.71, p = .007, d = -0.25, 95% CI = [-0.42, -0.07], BF10 = 3.49. When TO1 was presented in the same size as positive SOs, and TO2 was presented in the same size as negative SOs, participants preferred TO1 over TO2 (M = -7.56, SD = 57.83). When the size contingencies were reversed, participants preferred TO2 over TO1 (M = 7.06, SD = 61.09)."

In the more typical way of reporting such results (see my former review), they found a priming effect (showing the relative positivity of the initially neutral target that was paired with the positive stimulus) of M = (7.56+7.06)/2 = 7.31 ms; standard deviation is (assuming equal sample sizes) SD = squareroot(57.83\*\*2+61.09\*\*2)/2) = 59.5 ms. Thus, it is a priming effect of dz = .123. To evaluate this result, note:

-- According to the "rules of thumb" by Cohen (1988) already effects of dz = .20 are considered as small (.5 medium, .8 large).

-- An attempt to replicate dz = .123 with power 1-beta = 0.95 (.80) - alpha = .05 - would need a sample size of N = 861 (521).  
-- The meta-analysis by Herring et al. (2013) yielded an average effect size for evaluative priming (using the evaluative decision task) of d= .45.

**Authors**: In line with Reviewer 2’s suggestions, we have revised our paper to make it clear that the evaluative priming effect we obtained was small (see changes on p.43) and unpack potential reasons for this in the General Discussion (see changes on .p67).

**Reviewer 2**: The authors wrote in a footnote: "Given the unreliable nature of the evaluative priming effect, large sample sizes typically required to find EP effects," That is not entirely correct. To replicate dz = .45 with power 1-beta = 0.95 (.80) - alpha = .05 - one needs a sample size of N = 54 (33). Of course, the meta-analytic result was found by using a priori clear positive and negative primes. Thus, I would agree that the sequential Bayes approach is in principle a good idea. However, the outcome of dz=.123 is rather disappointing.  
  
[An aside: the word "unreliable" is misplaced here. It is correct that the EPT as an implicit measure of individual differences in attitudes (e.g., if black and white faces are used as primes to establish a measure of individual differences in racism) is unreliable (e.g., split-half correlation is very low). Here, however, we are not faced by a case of measuring individual differences. For each participant, there is designated positive and a designated negative stimulus. Thus, the "benchmark" is given by mean priming effects, as, e.g., reported by the meta-analysis and not by some evidence of unreliability.]

**Authors**: In line with Reviewer 2’s suggestion we have removed the following material from our manuscript (“Given the unreliable nature of the evaluative priming effect, large sample sizes typically required to find EP effects”) and simply refer to the sequential Bayesian testing approach adopted (see footnote 5 on p.40).  
  
**Reviewer 2**: (2) The authors mentioned the sequential testing procedure only in a footnote. In all other respects they report the results as if the experiment was conventionally planned. I do not think that this is ok. I would recommend that the sequential testing procedure is put more into the foreground and that the conventional t-test results are reported only as a kind of additional information (a kind of service for readers who are not accustomed to Bayes statistics.)

**Authors**: In line with Reviewer 2’s suggestions we have put the sequential testing procedure more in the foreground (see changes on p.40) and include t-test results as additional information (see changes on p.42).  
  
**Reviewer 2**: (3) By making the sequential testing more clearer, the authors must put more into the foreground that they used a very unconventional upper criterion of BF10 > 3 (that, without giving a reason, deviates from the lower criterion of BF01 > 6). For example, I cannot imagine that reviewers of a registered report would accept that authors plan with BF10 > 3 for a seq Bayes study.

**Authors**: We now put in the foreground that we used an upper criterion of BF10 > 3 (see changes on p.40).   
  
**Reviewer 2**: What follows? I do not know. Maybe it is enough to recommend that the authors are more clearly about the EPT results. For example, they should not insinuate that the low effect size has first of all its origin in the weakness of the EPT paradigm. The decrease from dz = .45 to dz = .12 has something to do with the present approach. It might have something to do with massively repeating two prime words. (Actually, I do not know what would happen if one would use only "love" and "war" in an EPT), but it might have something to do with transfer of valence due to a shared feature being not a very robust phenomenon.  
  
Thus, the report should not hide the weakness of the results such that any readers who are not convinced by the IAT results will think twice before s/he will conduct follow-up research on this new paradigm.

**Authors**: In line with Reviewer 2’s suggestion, in the General Discussion, we now acknowledge the small nature of the effect and offer several suggestions for why this might be the case (see changes on pp.43, 64).   
  
  
Reviewer #3: Signed: Yoav Bar-Anan  
  
The manuscript reports eight experiments that tested the effect of presenting a neutral stimulus that shares a feature with an affective stimulus on the evaluation of the neutral stimulus. In Experiments 1 and 6, the neutral stimulus and the affective stimulus changed color to the same color (the experiments differed in that three stimuli appeared in each trial of Experiment 1, and 8 in each trial of Experiment 6). In Experiments 2-3, the neutral stimulus and the affective stimulus remained in the same color and a third stimulus (the opposite valence) changed color. In Experiments 4 and 7, the neutral stimulus and the affective stimulus (both words) appeared in the same font size (Experiment 4 used the IAT as an indirect evaluation measure, and Experiment 7 used evaluative priming). In Experiment 5, the neutral stimulus and the affective stimulus appeared in different colors but previously the participants learned that the colors go together (in a match-to-sample procedure). In Experiment 8, each stimulus appeared alone on the screen, but the shared feature was that participants read that the neutral stimulus and the affective stimuli were pulled from the same bag. In all the experiments, excluding Experiment 2, the authors found that people preferred the stimulus that shared a feature with positive stimuli of the stimulus that shared a feature with negative stimuli, both when evaluation was measured directly and indirectly. The authors argue that this effect can explain previous findings in psychology.  
  
1. As in my review of the original submission, I believe that the revised manuscript has a very strong potential to make a positive contribution if shared with the rest of the scientific community. The revision is definitely stronger than the original submission. I am very eager to see this paper published because I think that it provides a very useful step forward in the understanding of the effect of pairing on evaluation. I have only a few comments the authors might consider, for further improvement of the manuscript's contribution.

**Authors**: We thank Reviewer 3 for his kind words and ongoing feedback, the latter of which has helped us to further improve our paper.

**Reviewer 3**: 2. I previously commented about inconsistency in the writing of the methods and the results, with some confusion on what TO1 and TO2 stands for. I worry that I must be missing something, but I still think there were inconsistencies. In Experiment 1's method (p. 1), the authors wrote there were "two types of trials: one trial in which Target Object 1 [TO1] was eventually presented in the same color as positive source objects, and another trial in which Target Object 2 [TO2] was eventually presented in the same color as negative source objects." And they added "Assignment of Morag or Struan to the role of TO1 and TO2 was randomly counterbalanced across participants." So, TO1 and TO2 seem to be "roles", with TO1 being the stimulus that shared features with positive stimuli and TO2 the stimulus that shared features with negative stimuli. That is consistent with the text that explained how scores were computed (p. 14): "a difference scored [sic] was calculated by subtracting scores for TO2 from TO1. Positive values indicate a relative preference for the TO that eventually shared a color with a positive SO over the TO that shared a color with a negative SO. Negative values indicate the opposite." But, in all the experiments, the results were a comparison between a condition in which TO1 was paired with positive and a condition in which TO1 was paired negative. For example, in p. 15: "When TO1 shared a color with a positive SO and TO2 shared a color with a negative SO, participants showed a relative preference for TO1 over TO2 (M = 3.33, SD = 4.60). When the color contingencies were reversed, participants preferred TO2 over TO1 (M = -4.15, SD = 4.48)."

**Authors**: After re-reading the paper again we see the potential source of confusion that the Reviewer is experiencing. We have revised the description of the experiments throughout the entire paper once more to avoid this issue. Specifically, we now omit any reference to the idea of TO1 and TO2. Instead we acknowledge that there are two target objects in each experiment (MORAG and STRAUN) and that assignment of MORAG and STRUAN to the ‘role’ of the target object related to positive or negative objects is counterbalanced across participants. Moreover, in the Results sections, we now say that “scores differed depending on the valence of the SO that shared a color with a TO. When a Target was presented in the same color as positive SOs, and a second Target was presented in the same color as negative SOs, participants preferred the former over the latter. When the color contingencies were reversed, participants preferred the latter over the former”. Hopefully this new phrasing avoids the confusion that the old phrasing elicited.

**Reviewer 3**: The results reported for Experiment 2 were further inconsistent and confusing (p. 18): "Participants generally showed an automatic evaluation favoring TO2 over TO1 regardless of whether (a) TO1 remained in the same color as positive words and the color of negative words changed (M = -0.26, SD = 0.54), or (b) TO1 remained in the same color as negative words and the color of positive words changed (M = -0.12, SD = 0.60)." But what is TO1 if it sometimes shared color with positive words and sometimes shared color with negative words? According to the discussion later, TO2 is the word that shared a color with negative words (p. 19): "Participants preferred a target when it shared a color with negative source objects more than when it shared a color with positive sources." So, what are (a) and (b) in the results reported earlier?

**Authors**: See our previous comment.

**Reviewer 3**: In Experiment 4, in the methods, the authors continued treating TO1 and TO2 as specific roles: "This time TO1 and positive sources were presented in the same sized font whereas negative sources were presented in a differently sized font. Likewise, TO2 and negative sources shared a common sized font whereas positive sources were always presented in a different sized font." But the results reported "When TO1 was presented in the same size font as a positive SO and TO2 was presented in the same size font as a negative SO, participants showed a relative automatic preference for TO1 over TO2 (M = 0.16, SD = 0.48). When the size contingencies were reversed, participants demonstrated a relative preference for TO2 over TO1 (M = -0.18, SD = 0.46)." But, didn't the authors say in the methods that TO1 is the stimulus that shared a feature with positive stimuli and TO2 is the stimulus that shared a feature with negative stimuli? As I said, I must be missing something very simple, but I can't figure out what it is. Maybe in the methods, TO1 and TO2 are examples and the authors mean that they counterbalanced the meaning of TO1 and TO2 between participants? If that is the case, it definitely confused me, so, for the very least, that information should be provided very explicitly, and perhaps more than once.

**Authors**: See our previous comment.   
  
**Reviewer 3**: 3. In the revised manuscript, the authors now provide more details about the data they collected about what the participants thought about the experiments, which is a big step forward, in comparison to the previous submission. However, that information could be presented more clearly. The readers of the main text would not have a clear understanding what the authors mean by each type of awareness (influence awareness, hypothesis awareness, and so on). The authors do not explain what these are in the main text. In the revision letter, they wrote that the precise wording of the exploratory questions is outlined in the pre-registered OSF files connected to this paper. However, the link they provided (https://osf.io/pqm9v/) leads to two compressed files that pertain only to the first experiment. In one of the zipped folders (Design.docx), wording for the questions appear, but this is definitely not easily accessible. I still think it would be helpful for readers to see in the main text at least one example for the wording, perhaps in another table, in order to understand what the authors mean when they discuss this data (before the General Discussion).

**Authors**: In line with Reviewer 3’s request, we now include the exact items that comprised the exploratory questions from Experiments 1-8 (see changes on p.14-15, 20, 23-24, 27-28, 32, 36-37, 48-49).

**Reviewer 3**: 4. In the materials I found (the Design.docx file), I was confused by the "hypothesis awareness" question: "During the first part of the study, did you notice that the color of MORAG and STRUAN switched to the same color as either positive or negative words? Please be honest here". Why does that question measure hypothesis awareness? I assumed hypothesis awareness is awareness about the hypothesized effect. Aren't the participants supposed to notice the shared feature (switching to the same color as either positive or negative words) in order to reach contingency awareness? In fact, the contingency awareness question is very similar: "In the first part of the experiment (when words appeared initially in white and then switched their color) did MORAG/STRUAN switch to" with the responses "The same color as positive words", "The same color as negative words", and "I don't remember". So, isn't it very likely that people who remember the contingency would also respond affirmatively to "hypothesis awareness" question?

**Authors**: We appreciate Reviewer 3’s general point: hypothesis awareness is a term that is usually employed when referring to a participant’s awareness of what the experimenter was trying to achieve. Yet looking back at the different ways we phrased our ‘hypothesis awareness’ question it becomes clear that what we were actually interested in (and measuring) was something else – namely – the extent to which the participant was aware of the *feature* shared by the target and source objects (e.g., color, size, location). Our original intention here was to see if participants were (a) aware of this shared feature and then (b) if they used this feature when generating their evaluation of the target object. Our ‘hypothesis awareness’ question was designed to target (a) and the influence awareness question (b).

Therefore, we have gone back throughout the manuscript and relabeled the ‘hypothesis awareness’ question as the ‘shared feature awareness’ question. The Reviewer’s comment also made us go back and consider if the other questions were correctly labelled and targeting what we thought they were targeting. This has led to several changes in how the items are now described. Critically, we have included a new footnote that highlights, and provides a rationale for, the deviation in labelling between the manuscript and the OSF files and included such an explanation on the OSF project page as well (see footnote 4 on p.16).

**Reviewer 3**: 5. I was also confused by the rationale behind excluding participants who showed awareness for the influence of the feature sharing on their evaluation. If the authors believe that inference is a reasonable account for their finding, why wouldn't participant be aware of the influence of the feature sharing on their evaluation? I was further confused by the text that addressed the exclusions. In p. 47, the authors first write that the results were robust to including only participants who were not influence aware. But, about 10 lines below that text, they write that the effect "was not found to be robust to requiring participants to be influence unaware." Table 2 did not help to reduce this confusion, because under the "Excluded subset" column, the authors wrote "unaware" in the "influence awareness" row. So, did they exclude those who were unaware? Probably not, because in that column they also wrote that people aware of the contingency were excluded. So, probably the column's label was supposed to be "Included subset". And, in Table 2, it would have been helpful to also see the effect size and statistical significance without the excluded subset (the best would be to report these details for both the excluded and the included subset).

**Authors**:

Regarding influence awareness, the Reviewer’s comment gave us pause for thought regarding our reasoning here, not only for the direction of the prediction but as to why we were including this analysis of influence awareness at all. In order to retrace our own thinking, we consulted our preregistrations. In doing so, we realized they referred only to contingency awareness, demand compliance, and reactivity (with one exception: study 7 did refer to influence awareness once).

Our primary stated goal here was to take known moderators of EC effects and examine whether this also applied to the shared features effect. Whereas contingency awareness, demand compliance, and reactivity are all moderators that are commonly considered in the EC literature, influence awareness is not. In order to align the manuscript with our pre-registered researcher questions and aims, we have therefore removed influence awareness from the robustness and moderation analyses.

Aside from our preregistered research questions, we think this is also justifiable on the basis of conciseness and clarity. The manuscript is now close to 80 pages, and exploratory questions regarding influence awareness would seem to detract from our key points here. Of course, the influence awareness data is freely available for others to use or meta analyze across this and other studies. Indeed, the reason for including this measure in the absence of clear hypotheses was to allow us or others to consider this variable at another point in time, outside of the current manuscript.

Regarding the clarity of the results from the moderation meta analyses, we have (a) reformulated Table 2 to report the effect sizes in each subgroup separately (rather than the difference between the subsets), and (b) more clearly labelled which subset is being referred to in each case. We thank the Reviewer for their input here and hope that this clarifies the matter for the reader.

**Reviewer 3**: 6. To be clear, my concern in this case is for the clarity of scientific communication. I am not much concerned about the possibility of demand effects. Not because it is an impossible account for the finding, but because it is difficult to refute that account. Further, the authors have done much toward this goal, considering the early stage of the investigation (the first paper to report or identify this effect), and the standards in our field. First, they measured evaluation with direct and indirect measures. Using indirect measures is helpful, although not definite because the memory that mediates the effect of demand on self-report can also influence indirect measures. In other words, if I remember that you want me to like Morag, that memory can influence the IAT and evaluative priming, even if I don't really like Morag. Second, Experiment 8 found that the effect persisted even when participants were told explicitly that there was no "connection between the words and images that are pulled from each bag". That finding is good initial evidence against the demand account (in the future, that evidence might be further improved by fixing the confusing statement that "the contents of each bag were randomly created". That statement might be confusing because the participants see later that positive words come from one bag and negative words come from another bag, which is clearly not random).

**Authors**: We see the reviewer’s point here: certain participants might be confused about whether the randomness refers to (a) the contingencies that they are seeing between the source and targets emerging from each bag (which are not random), or (b) the fact that the targets have been *randomly* assigned to the same bag as certain sources (which can be a perfectly random allocation). The instructions in Experiment 8 are ambiguous in that they could be referring to condition (a) or (b). We now acknowledge this in the General Discussion and highlight it as a direction for future study (see footnote 9 on p.64).   
  
**Reviewer 3**: 7. To clarify further, when I mentioned a communication effect in my previous review, I did not mean experimenter demand. By communication, I mean that the induced attitudes are quite real (unlike in a demand effect), but the cause for the formation of the attitudes is unique to the experimental settings. Specifically, perhaps the participants tended to assume that by pairing stimuli (through a shared feature) the authors are telling them that the stimuli are similar. In contrast, if people, in real life, tend to believe that red apples share more attributes with red leaves than with green leaves, it is probably not because they think that someone has painted the leaves red to convey to them that is shares other attributes with red apples. Both in real life and in the experiment, the effect might occur because people automatically infer similarity whenever they understand that stimuli share a feature (even when that inference makes no sense like in the experiment, or in the leaves/apples example), or because features overlap creates a link in memory between the stimuli, which later influences judgment and other evaluative reactions due to spread of activation (or another automatic retrieval process). In contrast, only in the experiment, a communication explanation is a reasonable alternative account.

The same challenge is also still valid for evaluative conditioning, an effect that has been studied in hundreds of studies, and has hardly ever addressed that challenge (it is much less valid for other forms of information that mimic reality more closely, like reading about a target's behaviors and traits). Therefore, I assume that it is an alternative account that most researchers consider improbable (unlike me), or one that is very difficult to test. The "at random" condition in Experiment 8 is a rare step toward facing this challenge. For that reason, I believe that the readers would benefit if the authors emphasize the importance of this condition more strongly, perhaps even in the abstract.

**Authors**: In line with Reviewer 3’s suggestion we now emphasize the importance of this condition more strongly than before and now make reference to it in the abstract as well (see changes in abstract and on p.9, and footnote 9 on p.64).  
  
**Reviewer 3**: 8. If I understand correctly, Experiment 2 did not find statistically significant evidence for the effect of the manipulation on evaluation. Yet, the authors interpret the results as evidence against their hypothesis and, in the GD, as possibly informative finding about moderators of their effect. Given the inconclusive evidence, more caution might improve accuracy.

**Authors**: We have revised the results and discussion section of Experiment 2 and the General Discussion in line with Reviewer 3’s suggestions. Specifically, we now state that we did not obtain evidence for a difference in evaluations as a function of the shared feature manipulation and deploy more caution when interpreting our findings (see changes on pp.21-22).  
  
**Reviewer 3**: 9. In footnote 5 and in the text in p. 44, the authors refer readers to https://osf.io/vbk54/. That link is currently private.

**Authors:** We thank the reviewer for highlighting this and have updated the OSF links throughout the paper and made them fully open access.

**Reviewer 3**: 10. In p. 13, the authors present the exploratory questions but do not indicate to the readers that they will cover the results of those only later, meta-analytically. It would improve clarity if the readers are informed about that in advance.

**Authors**: In line with Reviewer 3’s suggestion we now mention early in the paper that exploratory questions are only considered meta-analytically and not within each study (see changes on p.15).

**Reviewer 3**: 11. I am still doubtful about the strength of the authors' examples for known findings in psychology that are based on the shared-feature effect. I agree that EC (and attribute conditioning) seems like an instance of this effect, and I have already said that I believe that it is the main contribution of the present research (another important contribution is the clever insight about the similarity between acquisition and generalization). Other than EC, the most well-known and important topic of research mentioned by the authors pertains to the minimal group paradigm. However, to the best of my knowledge, the interesting effect in that paradigm is that the minimal-group manipulation is strong enough to cause pro-ingroup bias and discrimination, and not people's assumptions about similarity between group members (I would be surprised if even 10% of the countless papers that mention minimal group mention the similarity finding). I also do not share the authors' judgment that there is much similarity between the manipulation in Experiment 8 and the minimal group paradigm. I am not an expert in minimal-group, so my concern is not well-informed. Therefore, I only recommend to the authors to verify with an expert that the effect to which they refer (the assumption about similarity between group members after a minimal group manipulation) is not quite esoteric in its importance, in the context of minimal-group research.

**Authors**: The minimal groups paradigm refers to a *procedure* where a person is (randomly) allocated to a group based on some stimulus (e.g., flip of a coin), stimulus property (e.g., a person is wearing a similar piece of clothing as another person), or response (e.g., a person says that they like one type of painter over another). After this procedure the minimal groups *effect* emerges, such that the individual’s behavior changes towards members of the group they were assigned to, and members of other groups. This behavior can take many forms, and in most studied cases, researchers have focused on one set of outcomes (intergroup bias and discrimination).

Put simply, a minimal groups paradigm effect emerges when an individual is randomly assigned to one group and not another group on the basis of some stimulus, stimulus property, or response. In most cases, the individual being assigned to the group is the participant, and the outcome of interest is how that individual responds to other individuals (inside or outside their own group).

In Experiment 8 we also adopted a procedure where an individual is randomly assigned to a group on the basis of a stimulus (random allocation by the computer – much like the flip of a coin). As a result of this assignment behavior changes. Critically, however, there are several differences between the task we used and that typically used elsewhere in the literature: (a) the individual is no longer the participant, rather an unknown individual is, (b) the group contains valenced stimuli rather than other people, and (c) the outcome is not intergroup bias but changes in person perception.

So we see the Reviewer’s point: there are important differences between our task and that used elsewhere in the literature. One option would be to simply remove all mention of the minimal groups paradigm from the paper and label our procedure a ‘person perception’ task or something similar. On the other hand, we believe that retaining the notion of minimal groups is not only possible but also allows us to highlight new ways forward.

As far as we can see, no *a priori* reason why a minimal group paradigm *has* to involve a specific set of procedural parameters (e.g., participant as individual; group comprised of other people, focus on intergroup bias). Sure, these parameters have been used in the past. But that is not a strong justification for limiting what is used in the future. If anything, our shared features approach shows that there may be entirely new minimal group effects that have not been studied and which could open up interesting new avenues for exploration (i.e., minimal group effects where the individual is not the participant, where the group members are not necessarily other individuals, and the outcome is not only intergroup bias).

We have revised our paper with this latter idea in mind. We now acknowledge the differences between our procedure and those elsewhere in the minimal groups paradigm literature, and highlight the potential of looking at this effect from the perspective of our shared feature account (see footnote 6 on p.46).