Manuscript No. XGE-2019-1738  
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 feature principle: If two objects share a feature, people assume those objects also share other features” (XGE-2019-1738) that you recently submitted to Journal of Experimental Psychology: General (JEP:G). I was fortunate to receive comments and evaluations from individuals who are very knowledgeable and highly respected experts in the topical area you are investigating. As you will see when you read their critiques, the reviewers made a great job and offered many detailed points and constructive suggestions centered on improving the current paper.

I read the manuscript prior to receiving these reviews in order to gain an independent perspective on the paper, and then again with the reviews in hand. In the end, there turned out to be a considerable level of consensus among the majority of us with respect to the perceived strengths and limitations of the current paper. All of us found merits to your article. Reviewer 2 judged your paper interesting and thought-provoking, while Reviewer 1 found that your arguments hold intuitive appeal. In the same way, Reviewer 3 thinks that your study could help broaden the scope of research in this area, something that is expected from an article in JEP:G. Having read your paper, I share these positive assessments. Though being not an expert in this domain, I found this reading interesting and enjoyable. At the same time, all the Reviewers raised several concerns that are important enough to prevent a publication of the paper in its current form, and I must also say that I share most of them. Although Reviewers’ recommendations concerning publication almost span the entire range of the possible options, it is my opinion that your hypotheses and findings have the potential for a valuable contribution that would deserve publication in JEP:G.

Consequently that I would like to encourage you to submit a revision. Please note that this is not a guarantee your manuscript will be published.

I would ask that your revision centers on five main points

**Editor:**

1. The first point concerns your experiments. Although Reviewer 2 was impressed by your paradigm (I was too), Reviewer 1 notes that along with the way you describe your experimental set (one neutral CS shares a feature with a positive US, and the other neutral CS shares a feature with a negative US), another description is possible because one CS is always the odd one out. Thus, instead of a shared-feature principle, it could be an odd-one-out principle, creating a confound in your experiments. It might of course be assumed that detecting that there is an odd one out requires having detected the feature shared between the two other items, but it remains that it would be safer to provide evidence that the shared-feature principle works even when the odd-one-out principle cannot. Reviewer 1 suggests one way to circumvent this difficulty by using four stimuli. I think that this control experiment is needed, and I invite you to add it in your revision.

**Authors**:

**Editor:**

1. The second point concerns the use of the IAT. Although as I said above I am not an expert in this domain, I knew that the IAT task has been the object of criticisms, and I was not surprised to see that one of the reviewers (Reviewer 2) recalls this point. Maybe you have good reasons to think that these criticisms do not apply to your study, but it might be safer, and this would certainly strengthen your demonstration, if you used one of the tasks suggested by Reviewer (affective Simon, evaluative priming, recoding-free IAT).

**Authors**:

**Editor:**

1. Third, Reviewers 2 and 3 raise some questions about the way your methodology is described and your analyses presented. It is true that some details are missing and that our reviewers (and myself) had in some places difficulties to clearly figure out the details of your experimental design. This point should be relatively easy to address and would improve the readability (already excellent) of your work.

**Authors:**

**Editor:**

1. Fourth, both Reviewers 2 and 3 comment on your interpretation of the results. Both wonder if your effects might be based on participants interpreting the demand characteristics of the tasks or on some communication effects. I invite you to follow Reviewer 3’s advice to introduce in the main text the precise instructions given to participants (this is not wasting space, but providing what is probably the most important information that readers need: what were participants exactly asked to do?).

**Authors**:

**Editor:**

1. Five, both Reviewer 2 and 3 discuss your explanations at the functional and cognitive levels. I must say that I found them rather unsatisfactory myself. Reviewer 2 notes that your explanation about logical inferences sounds rather “old-fashioned” (instead of ACT Anderson’s model, I thought about L. J. Rips mental logic model, or Martin Braine’s approach, both models suggesting that there exists in mind a set of inferential rules triggered by available information to producing conclusions). I don’t think that being old-fashioned is necessarily something bad, certainly not, but both Reviewers suggest that the recourse to memory models could shed light on the reported phenomena. Without asking you to adopt a precise explanatory framework, you are of course free of your choices, I think that a more in-depth theoretical discussion would be welcome, especially for a JEP:G publication.

**Authors**:

**Editor:** If you decide to revise the work, please submit a list of changes or a rebuttal against each point which is being raised when you submit the revised manuscript.

What I expect is that you will take each of the concerns seriously and address these concerns in two ways. First, when possible, you should make changes in the manuscript to correct shortcomings that the reviewers perceive. (If there are comments that you do not find to be correct or apt, you still should consider that the incorrect perception is something that you might expect in other readers, so it would be helpful to take steps for the paper to anticipate such misperceptions and add clarifications in the text to prevent them.) The goal is to make your paper as accurate, scientifically responsible, interesting, and accessible to a wide range of experimental psychologists as possible.

The second way that I would like you to respond to the reviewers is to put a lot of effort into a careful cover letter that goes through the comments point by point, explaining how you addressed each comment and, if you disagree with a comment, why you disagree (and, if possible, how you altered the writing in anticipation that other readers might have similar concerns).

You might want to take advantage of a new policy as suggested in the editor’s interview on the journal’s  web page: “I also encourage authors to include a brief paragraph at the end of the article describing its broader context, explaining such things as how the ideas originated, how the findings are related to the authors' research program, and how the research will be extended in the future, much as one provides in oral presentations.” You may label it something like Context of the Research if you wish. By brief, I mean about the same length as an abstract, not more than about 250 words.

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-2019-1738 as invited by the action editor, Pierre Barrouillet.”

PLEASE NOTE:

As part of your Author Note, please provide the details (2-4 sentences) of prior dissemination of the ideas and data appearing in the manuscript (e.g., if some or all of the data and ideas in the manuscript were presented at a conference or meeting posted on a listserv, shared on a website, etc.).

Sincerely,

Pierre Barrouillet  
Associate Editor  
Journal of Experimental Psychology: General  
   
**Reviewer 1’s Comments**  
   
**Reviewer 1**: This paper argues that when two stimuli share one feature, people assume the stimuli also share other features. The authors demonstrate this with size, color, and transitive color relations as the shared feature, and positive vs. negative valence, as the assumed shared feature. I appreciate the transparency in the reporting of Experiments 2 and 3. Too often, papers present a clean, linear story that obfuscates the messiness of research. The argument presented holds intuitive appeal, but the following concerns make this paper less than convincing.  
  
**Reviewer 1**: 1. The empirical demonstrations are much narrower in scope than the paper's claims. Whereas the demonstrations are limited to size and (transitive) color and valence, the claims are discussed in terms of unqualified features. This mismatch between the framing and the experiments is quite jarring and I wonder if a more specialized journal in social cognition would be better venue. At the very least, the claims would need to be dialed back, and ideally, a much wider range of features (e.g., object preferences, social preferences) would need to be tested for this to appeal to a broad audience of psychologists.

**Authors**:   
  
**Reviewer 1**: 2. Across all five experiments, one neutral CS always shares a feature with a positive US, and the other neutral CS always shares a feature with a negative US. An equivalent description would be that one CS was always the odd one out. The authors privilege the interpretation of a shared features principle. But it could also be an odd-one-out principle: any stimulus that differs is evaluated in the opposite direction. Perhaps the authors can disambiguate these accounts by reanalyzing participants' explicit evaluations, but the relative nature of the IAT would make this difficult, if not impossible, without an additional control condition that can separate these out. Interestingly, the authors note this in the general discussion as a future direction, but it seems to be a confound in the present. The authors may counter that the salience alternative hypothesis tested in Experiments 2 and 3 addresses this concern. But the salience alternative concerns the change (at first all stimuli were the same color, but then they were not). The odd-one-out alternative isn't about change per se. Even without the change in color, there is still a stimulus that sticks out and that is what drives the effects - that stimulus becomes more different from the others in people's minds and the similarity between the stimuli that share a feature is constant. Using four stimuli instead of three and having two stimuli have something in common could be a way to disambiguate.

**Authors**:  
  
**Reviewer 1**: 3. The final experiment is the step in the right direction of demonstrating conceptual as opposed to mere perceptual similarity. But the manipulation of equating two different colors still falls within the perceptual realm and is therefore quite removed from the minimal groups- and moral spillover-type studies that motivated this extension. A more convincing case could be made through the following type of design. A neutral stimulus and positive stimulus are both sampled from the same bucket (the participant sees this happen), while another neutral stimuli and negative stimulus are both sampled from another bucket. If the shared feature principle holds, each neutral stimulus should acquire the valence of the bucket from which it was sampled. An even more convincing take would be if participants were told that the contents of each bucket were randomly determined and thus there were no such things as good and bad buckets, per se. This would mirror the minimal group paradigm whose flavor this experiment is trying to emulate.

**Authors**:   
  
**Reviewer 1**: 4. The authors' use of the term 'heuristic' in describing the value of the shared features principle is confusing. For many readers, a heuristic brings to mind a shortcut or rule of thumb that is often helpful but can sometimes lead to suboptimal choices. This doesn't fit with the authors' claim of organizing disparate phenomena across different parts of psychology.

**Authors**:   
  
**Reviewer 1**: 5. An interesting part of this paper is the suggestion that EC effects are driven not by spatial-temporal contiguity, as is often assumed, but rather shared features - which space and time are often among. If this is true, then EC-like effects should emerge even when space and time are not shared features. But in the current studies, space and time along with color are shared features. They aren't separated. Of course, the thrust of this paper isn't about EC per se but a more general phenomenon. But perhaps an expanded discussion of this in the paper's conclusion would address what would be on many readers' minds.

**Authors**:  
  
  
  
**Reviewer 2’s Comments:**

This is definitely an interesting, thought-provoking paper. It has three components.

1. The authors made the observation that a variety of (on first thought) diverse phenomena can be bundled under one abstract framing: It seems that if persons see that two objects share a feature they tend to judge the two objects as similar in other aspects (although it is clear from a "rational" perspective that the shared feature has nothing to do with the other aspects).
2. This observation leads to the development of a new paradigm (new, as far as I know) that is intriguing: a neutral target (a non-word) is presented together with a positive and a negative word. The neutral stimulus shares the color with either the positive or negative word and is evaluated accordingly in a later test phase (explicit rating and IAT).
3. The Interpretation of this phenomenon. The authors relate their results to own theoretical developments in learning psychology.  
     
   I find (a) a good point and this observation leads straightforwardly to (b); (b) I am impressed by the paradigm - its simplicity and its rather clear results. However, I have problems with the interpretation (c). I will discuss (a)/(b) ("The experiments") and (c) ("The interpretation") separately.

*The experiments*

As said, I found this new paradigm quite intriguing. There remain only two critical remarks that should be considered by the authors.

1. How much of the results is due to demand characteristics? That is, to what degree do participants think: "Oh, this non-word shares color with positive words; the experimenters want to tell me that I should consider this non-word to be positive as well. Well, why not doing them the favor!" Finally, I do not believe that this can be the dominant explanation for the results. But results might be a bit "contaminated" by demand. That should be discussed.

**Authors**:

**Reviewer 2**:

1. One way to argue against demand characteristics is to show change of valence with an indirect measure. But is the IAT the best choice (especially) here as an indirect assessment tool for valence? The IAT has been critizised for being burdened by recoding strategies (Rothermund & Wentura, 2004 [JEP:G]; Meissner & Rothermund, 2013 [JPSP]), that is, if target and attribute categories share a feature (not essentially valence), this feature might be used to solve the IAT task in the compatible block (with corresponding problems in the incompatible block). Here, we are faced by a situation that can be considered an exemplification of this argument: a non-word target shares an arbitrary, but salient feature (color) with, e.g., positive words. Thus, participants might use this feature overlap in the IAT. (I recognized that the color feature is no longer present during the IAT; nevertheless …). Moreover, whereas a demand strategy for the evaluation of the targets (see above) is eventually not very plausible, it is plausible that participants used feature overlap as a strategy for the IAT. Thus, especially since we are talking about a potential JEP:G-publication, it seems not too far-fetched to me to ask for a replication with an alternative indirect measure (affective Simon, evaluative priming, or recoding-free IAT).

**Authors**:

**Reviewer 2**:

1. A minor point with regard to the report of results: I found the report a bit "non-orthodox" in one aspect. The authors wrote (e.g., p. 14): "Self-reported ratings also differed as a function of whether the TO shared its color with a positive SO or a negative SO, t(98.32) = 8.33, p < .001, d = 1.65, 95% CI = [1.20, 2.10], BF10 > 10\*\*6. When TO1 shared a color with a positive SO and TO2 shared a color with a negative SO, participants showed a relative preference on the self-report ratings 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)."

TO1 and TO2 refer to the two nonwords that were used and which were of course counterbalanced with regard to the pos/neg assignment. Thus, essentially each participant has one nonword that was paired with positive sources and one nonword that was paired with negative sources. The straightforward result is that the mean within-participants difference in ratings is 3.74 [= mean of 3.33 and 4.15; see above]; it can be tested on deviation from zero in a one-sample t-test. (Outcome of inferential statistics remains almost the same.) Effect size then can be reported as dZ, which is the standard in cognitive experiments that are basically (i.e., disregarding counter-balancing) within-participants designs. Here effect size is - if I calculated correctly - dZ = 0.82.

(As indicated, the two ways of reporting the results are formally almost equivalent. This can, e.g., be seen in the power planning for a replication: dz=.82 leads to N= 22 [alpha=.05;1-beta=.95] for a within design; d= 1.65 leads to N=22 [11 + 11] for the two groups design. However, the report as a pure within-design fits better to standard cognitive experimentation.)

**Authors**:  
  
*The interpretation*

**Reviewer 2**: A preliminary note: Points #4 and #5 have more the character of a comment that - I am sorry - make clear that I found the theoretical interpretation not very appealing and fruitful; however, the interpretation is not wrong. But nevertheless: it appeared not new to me and the least to do is to carve out the older roots of this propositional account (e.g., the pure production system architecture of Andersons's early version of the ACT theory) such that it is clear for a reader that the authors return to theories that others might see as a bit "old fashioned".

The authors focus on two levels of explanation, one is called by them the "functional level". They exemplify by (see p. 32f): "To illustrate, imagine that you are presented with a positive word along with an unknown word. If this pair of stimuli is accompanied by the word 'SAME' this may signal to you that the unknown word has the same (evaluative) meaning as the positive word. As a result you will subsequently like the unknown word more than before. In this example the word SAME functions as a relational contextual cue: it signals that a relation of similarity exists between the unknown and positive word. One could conceptualize shared physical features such as color (Experiments 1-3) and size (Experiment 4) in much the same way: as a relational contextual cue which signaled a relation of similarity between two of the three stimuli presented in an acquisition trial (a neutral target and either a positive or negative source). Once such a relationship was formed other source features were transferred to the target (valence), thus leading to a change in evaluative responding."

I do not really understand the value of such a kind of explanation. The "SAME" example is merely an abbreviation for instructed learning of the vocabularies of a foreign language. That is, there is a rational person (i.e., this miraculous being whose causal "underpinnings" in terms of mechanisms we want to understand as psychologists) who understands an instruction. We never wondered about the capability of understanding an instruction at this level of description.

Another everyday example would be the relational cue of temporal contiguity: In the presence of a naïve observer, I press a light switch and the light goes on. If I do this several times, the naïve observer will grasp the relation. Thus, want the authors to point out that in this list of relational signals feature overlap is a further one?

If so, this reminds me (in the sense of an analogue) of the early Gestalt psychology in visual perception. There might have been a phase where an author could have claimed "Hey, I found another Gestalt 'law'!" That was of course commendable but did not eliminate the big problem of early Gestalt psychology: only description, no explanation. (I am a fan of Gestalt laws, but nevertheless they only set a starting point for an endeavor that might be only fruitfully fulfilled with a sound neuro-cognitive theory about the visual system.)

**Authors**:

**Reviewer 2**: The second level of explanation focuses on mental representations. The authors refer to a propositional logic of beliefs. There are two interpretations of this claim: First, this propositional logic is akin to our everyday psychology of personal beliefs which is still a psychology that focuses on the personal level (not on the cognitive or neural "underpininngs"). This theoretical language is the basis of making sense of ourselves and others. We take the "intentional stance" (Dennett, e.g., 1987 [Book]) towards ourselves and others and explain ("make sense of") our actions by recourse to beliefs and desires ("Why did he - after waiting for a quite a while - suddenly put 1000 $ on 'red' in the roulette? First, he wants to win money (desire); second, he believes that after 10 blacks, a red number must [almost necessarily] follow!")

This intentional stance psychology is of course indispensable for psychology because it is an explication of what we want to finally explain (i.e., we do not want to explain the mere physical behavior of putting some chips on the red number; we want to explain, e.g., what processes led to the false beliefs of statistical regularities). However, this kind of psychology does not deliver causal explanations (see Dennett, 1987). It is more a kind of "psycho-logic" (Smedslund, 1988 [Book]) that is not falsifiable by empirical research.

The second interpretation is that the authors address the cognitive level of explanation. That is, we can (provisionally) assume that the "machine" behind our actions is a symbol-processing computer whose algorithms work according to the laws of propositional logic. This is not wrong but in a sense very "old-fashioned". It reminds of the early days of Anderson's ACT theory. At the early stage the model was (as far as I remember) a pure production system, that is, a system built out of a large set of if-then rules and a working memory whose contents are step by step modified by applying the if-then rules. Later, Anderson made a hybrid system out of it by adding neural network like structures. The other reminiscence is of the project by Douglas Lenat who wants to create artificial intelligence by feeding a system with millions of propositions that capture any aspects of everyday knowledge. Has anyone heard of this project recently?

Moreover, it has to be proven whether such a "propositional machine" would stay consistent if propositions are added that are based on such a simple ("irrational") logic like "feature share -> transfer of other features")

**Authors**:

**Reviewer 2**: Isn't there an alternative interpretation in terms of recent accounts in basic cognitive psychology? I mean the event file framework (Hommel) or - admittedbly also a bit older - the object file framework by Kahneman (also: distractor-response bindings; e.g., Frings, Rothermund). Feature overlap might result in a binding of the target and the source. Thus, presenting the target retrieves the file including the source and its evaluation. This kind of theorizing has more potential ties to formal memory models and therefore to a level of explanation below the "person".

**Authors**:

**Reviewer 2**: I assume that the authors will reply that this kind of very basic theorizing cannot explain negative relationships (i.e., the UNEQUAL pendant to the SAME example mentioned above). But note, the present experiments do not address negative relationships. We have to believe the authors that their examples of negative relationships (published elsewhere) are comparable to the present results. Within the present manuscript, the alternative has to be discussed.  
  
**Authors**:   
  
Reviewer 3: Signed: Yoav Bar-Anan  
  
The manuscript reports five studies that tested the effect of co-occurrence of a neutral stimulus with two affective stimuli - one sharing a feature with the neutral stimulus that the other affective stimulus does not share that feature. In Experiment 1, participants observed in each of 48 trials three stimuli: a nonword, a negative word, and a positive word. After 3 seconds, the nonword and one of the affective stimuli changed color (the shared feature). One nonword always changed colors together with positive words, and the other nonword changed colors together with negative words. In a self-report evaluation questionnaire, in an IAT, and in a choice measure, participants showed preference for the nonword that changed colors together with positive words over the nonword that changed colors together with negative words.  
  
Experiment 2 was identical but in the trial sequence, one affective stimulus changed color and the other two stimuli remained in the same color (the shared feature). Participants were instructed to "pay close attention to the color of each word and how they change". This time, there probably was (the results were reported inconsistently; see later) a preference for the nonword that appeared on trials in which the positive word changed color over the nonword that appeared on trials in which the negative words changed color.  
  
Experiment 3 was identical, but this time participants were told that the nonword "will stay the same color as one of the words on the right. Please pay close attention to the colors of the words." This time, the preference between the two nonwords (in all three measures) changed such that participants favored the nonword that remained in the same color as positive words over the nonword that remained in the same color as negative words.  
  
In Experiment 4, in each of the 48 trials, one nonword appeared in the same font size as one of the affective word, and the other nonword appeared in a different font-size (either twice as large or twice as small). On all three measures, participants showed preference for the nonword that shared size with the positive word over the nonword that shared size with the negative word.  
  
In Experiment 5, participants first completed a match-to-sample task that paired two colors together and another two colors as a second pair. Then, in the 48 trials of the acquisition phase, the nonword, positive word and negative word appeared in different colors. However, the colors of the nonword and one of the affective words were colors that were previously taught to be a pair. In all three measures, the participants reported preference for the nonword that was paired in that way with positive words over the nonword that was paired in that way with negative words.  
  
In all the experiments, questions at the end attempted to gauge participants' memory, motivation and reasons for their behavior. The last three questions seemed the most relevant for measuring how the participants understood the experiment: (1) "Earlier you rated MORAG and STRUAN as being either positive, neutral, or negative. Did you base your ratings NOT on how you actually felt about those words but ONLY on what you thought the researchers wanted you to say?" with the responses "Yes", "No", "I don't know", (2) "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" followed by "Did this influence how you responded to question about your liking of MORAG and STRUAN? Please be honest here". Similar questions were used regarding size. In a meta-analytical analysis of Experiments 1, 3, 4, and 5, The authors reported that their results remained reliable even after including only participants whose responses indicated that they remembered the pairing but were unaware of an influence of the pairing on their evaluation.  
  
**Reviewer 3**: Point 1 - The manuscript reports a potentially important finding: people like stimuli that are paired with positive stimuli more than stimuli that are paired with negative stimuli, even when that pairing is not by co-occurrence. They also present a compelling conceptualization of that finding (people assume that stimuli that share one feature also share other features). I think that this manuscript would be a very positive contribution to the literature. In recent years, research on the effect of stimulus co-occurrence on evaluation (i.e., research on evaluative conditioning) has boomed and thinking on that effect has made much progress. Evaluative conditioning has become one of the two main forms of learning investigated by the researchers who study how likes and dislikes are formed (the other is verbal information about the target). In my view, the present research could help broaden the scope of research from a narrow focus on the effect of spatiotemporal co-occurrence to research on any form of pairing. As such, it is immensely important.  
  
I was also very impressed by the authors' open science practices. The materials posted online were very helpful. Even more commendable is the authors' decisions to report Experiment 2, despite the fact that its results were incompatible with the their initial hypotheses. It is very important to establish such reporting as the norm.  
  
I have a few suggestions that the authors might consider in a possible revision of this manuscript.

**Authors**:   
  
**Reviewer 3**: Point 2 - The introduction presents a very broad and general framing of the present effect. It is an inspiring framing, and I like it a lot. However, the experiments themselves are only about evaluation, and only about two specific features: color and size. Further, the stimuli paired by the shared feature always appear together on the screen. So, much research is still needed in order to generalize this effect: go beyond color and size, go beyond joined appearance at the same time, go beyond evaluation. It might serve the reader if the authors explicitly discuss in the General Discussion, the distance between the high-level abstract conclusion/hypothesis and the actual evidence they report in this paper.  
  
For the same reason, it might be more suitable to wait until the General Discussion with the broader conceptualization of the effect, beyond evaluation, and with the discussion about the possible application of this broader conceptualization for the understanding of already known effects.

**Authors**:   
  
**Reviewer 3**: Point 3 - One intriguing question about the present finding is whether it was an effect of communication. Did the participants understand the pairing in the present experiments as a message from the experimenter about the nonwords? Did the participants treat the questionnaires as tasks with correct answers? The communication account might be considered an effect of an experimenter demand, or it might be considered an informative finding that people understand other people's decisions to construct stimuli that share a feature as a symbol of similarity between these stimuli. That distinction is difficult to do, conceptually, and should probably wait for more evidence to accumulate. Yet, the present paper can do a better job in providing easy access to information that could help readers get a better sense of the experience of the participants in the study. That would help readers judge how likely it is that the participants understood the whole procedure as a form of communication constructed by the researchers to inform them about the valence of the nonwords. It is essential to provide more information about the wording of the instructions and the questions. I recommend summarizing most of them in the main text, and providing exact quotes of the most crucial parts. If that seems like too much information for the main text, please consider using a table, a figure, or at least a very explicit online supplementary document (not only the script used to program the study), with a link to that document in the main text.

**Authors**:  
  
**Reviewer 3**: Point - 4. As I said, the wording of the self-reported evaluation questions are important for evaluating this research. In the materials posted online, I was glad to see that the wording targeted the participants' feelings ("Please indicate how you feel about this item"). That reduces, to some extent, the likelihood of the possibility that participants used those questions to inform the researchers that they understood their message about the valence of the nonwords. As mentioned above, I recommend the authors to share that information more clearly with the readers.

**Authors**:   
  
**Reviewer 3**: Point - 5. Closely related, I was also very interested in the follow-up exploratory questions about the participants' experience in the experiments. Most interesting to me were a pair of questions: the first asked participants whether they noticed the pairing (i.e., what stimuli shared the manipulated feature), and the second asked them whether that pairing influenced their rating of the nonwords. I think that it would be informative to indicate the distribution of the responses to the second question in all the studies. I had a glance at the relevant data files posted online by the authors, and I estimate that about 30%-50% of the participants reported such influence. That made sense to me, considering the quite large effects reported by the authors. I usually see such large effects only in evaluative learning studies that provide very clear information about traits and behaviors of people. It seems that many participants understood the pairing as clear information about the valence of the nonwords.  
  
It was very informative to read that the finding survived the removal of those participants. That is great. I am not sure that many evaluative conditioning studies would survive the same exclusion rule. The details pertaining to those questions, the responses to them, and how the effect is moderated by the responses to those questions seem quite important to report in the main text (e.g., in a comprehensive summary section before the General Discussion), or at least in an online supplementary summary, with the link of the summary document included in the main text.

**Authors**:  
  
**Reviewer 3**: Point - 6. I accept the authors' decision not to focus, at this time, on a mental mechanism (or, indeed, on any mediator) of the effect that they report. However, I am not sure I follow their logic in the one mechanism they speculate about in the General Discussion - the Inferential reasoning account. As an example for that account, the authors suggest that people make the inference that "the source is good and therefore the target is also good" from the proposition "the target and source are similar in that they are both green". It is very likely that this kind of inference often causes that effect in everyday life. But, in the present experiments, the participants know that the researchers chose the colors for the stimuli. So, why would they make that inference? For that, the participants would need another proposition, for instance, "the researchers chose to present Morag in the same color as positive words because they know that Morag is positive". In other words, I find it difficult to understand how inference would explain the finding without turning it into a communication effect.  
  
Some other accounts do not suffer from this problem, and could easily apply to the effect in the present experiments and in everyday life. For instance, some of the shared feature manipulations might have caused perceptual grouping of the stimuli, linking stimuli of the same color (or the same size) together in memory. Later, when participants considered their liking of the nonwords, the association with the valenced stimuli was activated and was experienced as a gut reaction. That gut reaction influenced judgment. That account would also work without the perceptual grouping mediator at the encoding stage. From my perspective, any link in memory, even one that has been formed from the encoding of the proposition "this nonword has the same color as positive words" might cause this effect on evaluation, with no mediation of an inference from the proposition.

**Authors**:  
  
**Reviewer 3**: Point - 7. The description of the methods and the results was sometimes confusing. In p. 11, the text indicates that Target Object 1 (TO1) is the nonword Morag and Target Object 2 (TO2) is Struan, and that TO1 was paired with positive words whereas TO2 was paired with negative words. Then, in describing the IAT, in blocks 3 and 4, TO1 always shared a key (the left key) with positive words and TO2 always shared a key (the right key) with negative words. From that description, it appears that there was no counterbalancing of the nonwords to a role (whether they were paired with negative or with positive words). The is highly unlikely, so, I assume that Morag was sometimes paired with negative stimuli, and sometimes with positive stimuli. But then, in the IAT, was it the case that Morag always shared a key with positive words in Block 3? Or perhaps the nonword that was paired with positive words in the acquisition always shared a key with positive words in Block 3? The appropriate design would be to counterbalance everything: which nonword was paired with positive words, which nonword shared a key with positive words in Block 3, and whether the nonword that shared a key with positive words in Block 3 was the nonword that was previously paired with positive words. I assume that's what the authors have done, but it was not clear to me from the text. Therefore, I recommend the authors to improve the clarity of the methods description and to indicate explicitly and clearly what counterbalancing they have done in the experiments.

**Authors**:  
  
**Reviewer 3**: Point - 8. More confusion occurred in p. 17. The authors reported comparing two conditions: (1) TO1 remained in the same color as positive words and the color of negative words changed; (2) TO2 remained in the same color as negative words and the color of positive words changed. I thought these were the two trial types in the acquisition, not two conditions manipulated between participants. The two conditions specified in the test of the self-reported evaluation in p. 18 seem more compatible with the description of the task. But, then in the description of the results of the behavioral intentions in p. 18, the authors report a preference for TO2 over TO1, with no specification of a condition. Considering how they later interpret the results, it seems they now use TO2 as the name for the nonword that always remained in the acquisition in the same color as the negative words. In the results of Experiment 3, which was almost completely identical to Experiment 2, the results are reported more consistently: TO1 and TO2 refer to the nonword stimuli and change their role between participants (sometimes TO1 remained in the same color as positive words and sometimes in the same color as negative words).  
  
**Authors**:

**Reviewer 3**: Point - 9. The question that the researchers used to measure experimenter demand might have been confusing: "Earlier you rated MORAG and STRUAN as being either positive, neutral, or negative. Did you base your ratings NOT on how you actually felt about those words but ONLY on what you thought the researchers wanted you to say?" with the responses "Yes", "No", "I don't know". This wording is confusing because of the use of negation and the very short answers (the participants had to affirm NOT doing one thing but ONLY another thing), and because there are a few different statements to affirm or negate. Should I respond "No" if I think I have changed my feelings about the nonword based on what I thought the researchers wanted me to feel? Perhaps it would exhaust the readers to read a discussion about the challenges of wording questions about self-perception of what the participants did in the experiment, so the authors were probably correct not to discuss it. Still, it is another reason why improving the readers' access to the wording of the questions (which I found only by looking at the scripts used to program studies, posted online) might help the readers evaluate this work better.  
  
**Authors**:

**Reviewer 3**: Point - 10. The importance of the IAT results is that they suggests that the pairing might have influenced automatic evaluation (as the authors note in in p. 9), and not only responses to the self-report questions. However, it is currently unclear that effects of pairing manipulations on the IAT reflect an effect on automatic evaluation. The IAT is probably the best tool we currently have to test that, and I use it often in my research for the same purpose. Nevertheless, it might help readers if the authors remind them that the IAT results in the present research provide only initial evidence that the manipulation influenced automatic evaluation.  
  
**Authors**:

**Reviewer 3**: Point - 11. In recent years, most evaluative conditioning researchers use De Houwer's definition for that effect: a change of the evaluation of the CS as a result of pairing with a US. Although defining EC as an effect had a very positive influence on the field, the word "pairing" might have been a problematic choice because EC research has always been about co-occurrence and not about all forms of pairing. I think that the concept "pairing" overlaps greatly with sharing a feature (all the present experiments paired stimuli by manipulating the sharing of a feature). Therefore, De Houwer's definition of EC actually fits better the present finding than it fits all the EC studies conducted so far, which were limited to pairing by co-occurrence (we noted that recently in Footnote 2 in Bar-Anan & Moran, 2018, and in Bar-Anan & Balas, 2018).  
  
In the present manuscript, I think this issue might cause some confusion. In pp. 7-8, the authors wrote that EC "is typically defined as a change in liking due to stimulus pairings. Most researchers would argue that EC effects are driven by the fact that CS and US are presented in spatio-temporal contiguity. Yet our account takes a different perspective. It argues that EC effects may actually be due to the fact that the CS and US share a feature with one another, and in EC studies, this shared feature just so happens to be the time and location at which they are presented. If correct, then the crucial element in EC is the fact that stimuli share a feature and not the mere fact that they are paired in space and time." It is odd because the definition of EC does not refer to pairing only in space and time, and does not limit EC to co-occurrence. From this section, it was not clear to me whether the authors think that the present experiments are EC effects. If not, then they probably think that the "pairing" in De Houwer's definition of EC pertains only to co-occurrence. If yes, then it would be helpful if they write that explicitly.

**Authors**: