Reviewer 1

Reviewer #1: You have done a careful job responding to the issues raised by the reviewers. Your revision provides initial evidence that choices between naturalistic stimuli are not altered in any significant way despite the presence of carefully calibrated choice sets. It may be the case that the attraction effect is very small for naturalistic stimuli, but it is possible that you will find non-zero effects for various important moderators. I will propose a reanalysis that avoids the confound generated by successive choices on the same AB pair, and an exploration of a number of interesting potential moderators.  
  
1. Clarify what you mean by naturalistic stimuli. There are at least two ways that choices among naturalistic attributes are fundamentally different from numeric or geometric ones. First, naturalistic stimuli, such as a city, a bran, or image of a person, have unique identities defined by many attributes. Second, those attributes generally are not monotonically related to preferences. Thus, a moderately scary movie can be preferred to one that is not scary or very scary. These two properties imply that it is difficult to find stimuli pairs about which one is unambiguously indifferent or strongly dominant. The cleverness of the paper arises because you limit that uncertainty by selecting for each respondent AB movies that get the same rating and Aa pairs with strong rating differences. It is important, however, to clarify the meaning and generalizability of the naturalistic stimuli you explore.

We thank the reviewer for raising this very important point. We clarified that naturalistic stimuli refers to options represented in a pictorial format as opposed to a numerical attribute by alternative format (page 2). In addition, we added the reviewer’s points highlighting the complexities around designing a choice experiment with naturalistic options in the beginning of the method section (page 4). **Tim, Neil, are you happy with this/shall I add to main text that this was pointed out by reviewer 2?**

*“Choice options with binary attribute dimensions (perceptual or numerical) can be relatively easily manipulated within a choice experiment, however, most real-world choices involve complex, naturalistic objects with a large number of underlying attributes.*

*These naturalistic options are represented in a pictorial format as opposed to by a set of numerical attributes.”*

*“When the stimuli have numerical attributes, it is straightforward to construct choice triplets with a target, competitor and decoy option. However, with naturalistic stimuli, this task is significantly more complicated. First, such stimuli can have a high number of attribute dimensions that might vary across individuals. In addition, it is entirely possible that preferences are not monotonic over these attribute dimensions (while this could also be the case for alternative-by-attribute representations, these dimensions are usually constructed to ensure monotonic preferences, e.g., probability of winning, amount to win).”*

2. Acknowledge the impact of past judgments within an MTurk survey. MTurk workers can lose compensation or credibility if they are inconsistent. Thus, in your study only 5% of choices select a decoy over the target which the respondent rated as at least 3 point less. That 5% rate is consistent with most survey takers striving to be consistent. The more important finding is that only 8.5% of the AB movie choices shifted in response to a different decoy. A consistent second AB choice is easy for respondents because the same AB choice had been made just a few minutes earlier. The problem, mentioned in the last review, is that the attraction effect measure must be zero for 91% of the choices where the initial AB choice repeats. A positive way to look at that result is 91% consistently ignored the decoys. A more likely account is that the MTurk respondent simply wanted to be consistent with that previous response. The previous selection violates the Huber et al. (2018) condition that there is no strong prior preferences in the second choice. It is a preference for consistency.

3. Redo the analysis ignoring the second choice. Since the current analyses confound a consistency effect with an attraction effect it is important to focus on the first choice. You show that overall .49 chose the decoy on the first choice. I am willing to accept that null result given the fact that AB pairs have equal ratings for each respondent. However, the moderation tests using both orders need to be replicated using only the first choice to appropriately uncover moderators for the attraction effect.

4. Test a number of moderators only using first choice. In Table 1 you test the TC similarity, TD similarity and TD rating difference across both choices and find no significant effects. Because 91% of the observations by construction have zero attraction effect these null results are just what one would expect. Repeat these analysis to using only the first choice. At the very least, I expect the range of responses in Figure 5 would show greater spread away from 0.5. Consider testing other variables. Does the degree of positivity (4-6) of the matched AB rating increase attraction? Does random order matter, so that having the competitor in the middle decreases the attraction effect? How about the effect of task order; does completing more tasks facilitate ignoring the irrelevant decoy? Are the results different for those who see many tasks? Finally, test the exclusion of respondents to see if respondents who take the test quickly or inconsistently are less prone to the attraction effect.

**Tim/Neil – Do you think it’s okay if I don’t add stuff about preference for consistency (“acknowledge the impact”)?**

We thank the reviewer for pointing out the importance of demonstrating the effect of various moderators on the attraction effect on the set of first choices. To investigate this, we repeated the analyses presented in Table 1 on the first choices only, and also ran a mixed effects logistic regression where in addition to the variables presented in Model 2 in Table 1, we added further controls, including overall preference rating for the target (and competitor); display order of the target, competitor, and decoy; overall number of choices completed; and current trial number. We believe that the stringent set of exclusion criteria we applied successfully filtered out participants who were not taking the task sufficiently **seriously (I am a bit confused by what she/he wants in last sentence – do you think this is appropriate reply?).**

*“Using only the first choices, we repeated the analyses presented in Table 1 (see Models 2 and 3 in Table A1 in the Appendix), and also tested the effect of various additional controls, including familiarity with the movies; target-competitor and target-decoy similarity ratings; target-decoy preference rating difference; target preference rating; display order of target, competitor, and decoy; overall number of choices; and current trial number (Model 5). When all covariates are included, we find some evidence that the larger the preference difference between the target and decoy, the less likely that the target will be chosen. Using Model 5, we estimate that the overall probability of choosing the target is .50, 95% CI [.39--.59].”*

Reviewer 3

Let me begin by saying I (still) like the paper and (unsurprisingly) still believe its conclusions. But I still have many objections. None of them are big, and most of them can be readily remedied.

**Abstract**: I think it is weird to say reverse the preference for the original two options. Context can, at the individual level, affect someone’s preference, but since you are mostly analyzing the effect of context for single decisions, it is weird to call this a “reversal.” At the aggregate level the decoy might change which of the two options is the modal choice, of course, but, again, though I think it is confusing to call this a “reversal.”

Later, you write “…found no evidence for the attraction effect in choice context where options have no numerical attribute dimensions.”

Regarding this, one tiny thing. I’d replace **have no** with **lack**. Possessing an absence is linguistically odd. More importantly. It is not about having or not having. I presume that nearly any attribute or dimension can be presented in numeric or non-numeric format. This is not a property OF the attribute or dimension but about the experimenter’s choice of representation. In some cases (e.g. probability) a numeric presentation is more natural. In other cases, (e.g., quality of view) a non-numeric presentation is more natural, and more informative.

**p. 1)** I don’t mind the cheesecake, pecan pie, and apple pie example, but please clarify whether this is a fictitious example (hypothetical results) or an actual example. I haven’t read Tsetos et al. so I don’t know.

I disagree with your claim that the attraction effect implies that preferences cannot be represented on a cardinal utility scale. Just use the example above. Ask about willingness to pay for each slice of pie with the lousy apple pie present or not. Suppose that affects WTP for the pecan pie. How is this a violation of cardinality? Nor do I agree with your claim that the attraction effect violates regularity. This only sometimes occurs. Retaining your example above, suppose the choice share of the desserts goes from 55, 45 to 40, 44, 16. I think most would consider this a (weak) example of the attraction effect, even though there is no violation of regularity. In my opinion, only the *final* sentence in this paragraph is accurate.

**p.2)** I find it bizarre when you write “*a natural concern is .. whether this hugely influential decision bias generalizes to real-world choice situations, where attributes often cannot be easily visually represented and compared*.” Huh? Can you provide an example? As you know, I think the issue is precisely the **reverse**. When I considered apartments, I could actually look out the window to appraise the quality of view. None shuttered with the number “7” affixed to the opaque area obscuring the view to communicate the “level” of that dimension.

Later you write “choice options with binary attribute dimensions (perceptual or numerical) can be relatively easily manipulated within a choice experiment.” This is miswritten. A binary dimension is something like sex which can only have two values (yeah yeah, go ahead and say there are actually 117 each with a unique pronoun and I’m a bad person who will go to hell). What I think you mean to write is that examples are simplest in choice situations that involve only 2 dimensions, like say, Octane rating and price among the gasoline options available at the pump. But both octane rating and price are *continuous* dimensions, not *binary* dimensions.

**p. 3)** Though I very much approve of what you did, and acknowledge that our procedures were more casual, I again want to point out that the Huber et al. list is regarding “things to avoid” to conduct a “real” test of the attraction effect is never really explained or justified: by you or them. I return to this objection at the end of my review. But, for now, let me list and discuss each.

(1) Strong prior preferences. As a heterosexual male, if I were presented with the option of having sex with an attractive woman or an attractive man, I’d choose the woman. And I very much doubt that preference could be altered by adding an average looking man to the choice set. Accordingly, while I certainly agree it would be hard to alter preferences that are very strong, this design is still a fully legitimate test of whether you could obtain an attraction effect in this context. To use your words, this *is* still an “attraction effect type choice scenario.” It reveals that contextual effects aren’t super strong. That isn’t surprising, of course, but it remains a fully legitimate test of the scope or strength of the effect.

Furthermore, with respect to the stimuli critiqued by Huber et al., where is the evidence that individual preferences were “too strong.” Speaking personally, I have a weak preference between apples and oranges. Where is the evidence that this is atypical?

Of course, weak preferences can be assured through various experimental contrivances. You do it in an unobjectionable way by creating bespoke pairs and triplets from a larger set of evaluated stimuli. In most of the prior work that I had critiqued in Frederick, Lee, & Baskin (2014), the weak preferences are created differently. If forced to choose between two flux capacitors whose fluon and bexor indices are either {6,5} or {4,7}, I’ll definitely have weak preferences because I don’t even know what you are talking about. And perhaps you can shift my choice from one to the other by adding some inferior third capacitor. But who cares?

(2) Inability to identify the dominance relation between the target and the decoy. Are you serious? Do you think that anybody given a choice between an apple, an orange, and a moldy orange would have any trouble identifying that an orange with mold is inferior to an orange without mold? To simply reflexively accept this “critique” without even providing a credible example is objectionable. Given the myriad stimuli used by us and by Yang and Lynn, you probably *can* find credible examples of this, but I’m not at all persuaded that this is a serious issue, in general.

(3) Heterogeneity of prior preferences between the target and the competitor. Huh? This is something to **avoid**? Huh? So I should only examine cases in which *everybody* chooses either A or B? I have no idea what you are attempting to say here.

(4) An undesirable decoy. Huh? As per your point (2) above, the decoy does, of course, *have* to be less attractive than the target, and, therefore, at least *relatively* undesirable. Also, under whose fiat is this an essential condition for a test of the effect? Retaining the example above, I’ve eaten citrus fruits with small bits of mold and thrown out those with more, and I suppose there is a dividing line where I’d be torn. But with respect to the effect of the presence of this flawed orange on my choice between an orange or apple, what is your theory exactly? Again, you are just parroting an unexplained objection cited by someone else, rather than *explaining* what you mean by “undesirable” and why this is an important thing to avoid.

(5) A decoy that is too desirable. In conjunction with (4) above, I’m now really in a bind, it appears, but I suppose I could thread the needle (maybe an orange with the tiniest imperfection in the peel). This is of course more of an issue for naturalistic stimuli than with highly stylized stimuli. A few people might legitimately prefer a view of a swamp than a view of a mountain, whereas nobody will prefer a view that is rated a “5” over one rated a “7.” The severity of this putative problem can be assessed by examining the data themselves. Obviously, if many people are choosing the decoy, then they are either choosing randomly or the experimenter has not successfully created the intended situation (or some combination). This potential problem is sometimes an actual problem but often not. Yet for some reason, the critics of our paper either couldn’t understand that or chose not to acknowledge it.

**p. 4)** Again, I don’t think that options “have” numeric attributes. I think that experimental stimuli represent attributes in a numeric or non-numeric way.

By the way, I’d prefer numeric or over numerical.

**Testing the attraction effect with real-world stimuli.**

This may seem petty, but again, *we* already did this, and did it in multiple domains. As have *others*. This title seems to imply otherwise. You did it better than others and should be proud of your procedure, which I think is excellent. But you should just label this “Testing the attraction effect using popular movies” or maybe “A carefully controlled test of the attraction effect using movies” “A test of the attraction effect using a sophisticated longitudinal procedure” or something like that. You are far from the first to use real world stimuli, but might well be the first to implement such a carefully controlled multi stage procedure, which I’m really impressed with.

I think you should delete the 2nd sentence of this section. Everybody will accept this. You don’t need to say it.

The movie stimuli in our paper did not necessarily involve distinctly different *genres*. As shown on p. 505, the sequels were obviously in the *same* genre as the target.

What do you mean by “identity role”?

**p. 5)** Figure 1 is just exceptional. It is super clear what you are doing, and these are excellent examples.

However, I think it is weird to say that decision makers can “identify” the inferiority of the decoy. Their ratings *define* what the decoy is. This is an issue again on the top of page 12.

**p. 9)** It isn’t at all clear to me why you are trying to *avoid* genre overlap between the target and decoy. I’d think this would be a *desirable* feature. Indeed, on bottom of page 12 you note that you created triplets so that the target decoy pairs would be as similar as possible. That makes sense. But then why were you trying to avoid genre overlap. I’m totally confused.

Regarding the question “How do you personally rate this movie?” your scale: 1(worst) to 7 (best) seems very odd, as it seems to be requesting a *ranking* not a *rating*. Worst of what? Best of what? All movies? Movies of this genre? Movies I’ve seen this year? What if it is the worst movie of a narrow genre, but still excellent?

On the same note, the idea of a preference *rating* also seems odd. Preference implies multiple options whereas a rating seems to involve an appraisal of a single option. I can like pea soup, without necessarily preferring it to chicken tenders.

**p. 10)** This is another excellent, clear example. However, it wasn’t clear exactly what respondents *did* here. Did the movie poster icons also function as choice options, such that if they clicked the mouse it indicated their choice?

**p. 13)** With respect to Figure 6, I once again find your nomenclature appended to the response scale highly odd, as subjects produced similarity ratings from 1 (least similar) to 7 (most similar). Why are you using *comparative* rather than *absolute* language here? I think that Terminator and Terminator 2 were similar and that both were very dissimilar from Driving Miss Daisy. I can make such judgments without considering any *other* movie. Indeed, I can barely understand what judgment you are actually asking subjects to make. I reckon I’ve seen about 2000 movies. Am I supposed to assess conceptual distance for the provided pair and rank it with respect to conceptual distance of the other million pairs of movies I’ve seen – e.g., Is *Terminator* is more similar to *Driving Miss Daisy* than *Hostel* is to *American Beauty*? What movie pair defines a “1”? What does a 3 mean? A 5?

Furthermore, assuming subjects interpret this comparative scale more like I did in the first part of the preceding paragraph than the second part, your design clearly achieved your goal of having more similarity between targets and decoys than between targets and competitors. That said, a considerable fraction of your target-decoy pairs were judged as not being very similar. Given this was a stated goal, did you consider eliminating trials where this obtained?

**p. 14)** When you wrote: “*Out of these 84 cases when participants switched, 48 times they chose the target both times and 36 times they chose the competitor both times*.” I momentarily thought: Huh, what does it mean to “switch”? But then I realized that the target at time 1 and time 2 were different movies; so all is good, but just note that readers might get confused here.

**p. 15)** I’d relegate much of this to an Appendix.

**p. 16)** Typo in line 2: While we have not **find** any…

Also, I wouldn’t say “…, *participants remained perfectly indifferent between the target and the competitor*.” That implies indifference at the *individual* level, and you forced each participant to choose. I’d just delete that phrase. You already had correctly stated the result (or non-result) with the words that precede it.

You write “*Our experiment is the first investigation to rigorously test the attraction effect with naturalistic stimuli whilst avoiding the five critical conditions set forth by Huber et al. (2014*).” If the conditions were “critical” why did you avoid them? I know what you are trying to say here, but it is very awkwardly phrased. More importantly, I reject that these conditions are, in any sense, critical, and as you never explain why they allegedly are. You just say “Huber et al.” Unfortunately for them and you, this isn’t very helpful – indeed, not even to those who have actually read Huber et al.

**p. 17)** When you write that choices have to be “constructed on the spot” what is the null? Do you usually choose what you are going to order *before* you see the menu?

Why *would* strength of the attraction effect be influenced by genre preferences? I guess it is fine to note that, but at most relegate it to a footnote.

You write “by creating bespoke triplets based on the ratings, we avoided individual heterogeneity in preferences as a potential confound.” Can you explain to me how heterogeneity in preferences is a *confound*?

\*\*\*\*\*\*\*\*\*\*\*\*\*

In conclusion, I like your study a lot. I like the stimuli and I like the design. And with many edits, this paper should be quite good.

That said, I think you made the paper worse by giving unwarranted deference to the 5 properties Huber et al. posit as critical. Suppose they had instead published a paper alleging: **(a)** You need to avoid stimuli that are red, **(b)** the decoy can’t be made of cheese, **(c)** The target and the competitor must be chosen in equal proportion among people from Oregon, **(d)** The chosen option should never be more than 8% better than the rejected option, and **(e)** In N-dimensional attribute space, the target must be at least 35% closer to decoy than to the competitor.

Accordingly, you conduct a study which observes properties **(a)** through **(e)** yet still find no appreciable attraction effect. I suppose this allows you to counter their implicit claim that the effect *would* be found in this subspace. You could write those three men and say “Ha!” While I, personally, might find that slightly satisfying, without *explaining* these five criteria, many other readers will be left confused.