There are a number of novel and valuable characteristics in this study. It raises the interesting challenge of assessing whether the attraction effect for naturalistic stimuli occurs if the requirements in Huber et al (2017) are satisfied. The authors' selection process for appropriate groups of test movies is outstanding. The researchers are able to identify AB pairs with ratings more than 3 on a 7-point scale that are equally rated by each individual. The A', B' decoys are similar to their A and B targets but have ratings at least 3 units less. Finally, the authors also do a good job screening inappropriate subjects whose ratings were performed too quickly, with too much breadth, or having too extreme autocorrelation with previous choices. It would be interesting to test whether the results change given various levels of inappropriate responses.

We thank the reviewer for their comments. To address whether our preregistered exclusion of inappropriate responses had a role in our core result, we repeated the central one-sample t-test on the raw dataset (no exclusion criteria applied) to test if the proportion of trials where the target was chosen was significantly higher than 0.5, but we found no evidence for the attraction effect in this dataset either, M = .5, $t(151)$ = -.72, p = .764, 95% CI [.48; .51].

We agree with the reviewer that the degree to which the apparent strength of the attraction effect can be driven/attributed to the inclusion/exclusion of inappropriate or low effort responses. However, due to the small proportion of trials that are excluded here, and the overall high quality of the responses in this particular dataset, we do not have sufficient statistical power to investigate this further.

The results are very consistent, perhaps too much so. Overall, the choices between A and B are almost equal, despite having seen the same A B choice in the context of different A' or B' decoys. It is unfortunate that the tests of similarity and familiarity have no effect, given other researchers have found that these variables moderate the attraction effect. It is important that that less than 5% choose the decoy, showing that choices are largely consistent with large differences in ratings. However, for those who do choose the decoy, a similarity effect could draw shares from the target, thus limiting the attraction effect. Thus, it is important to demonstrate that the results do not differ if the authors drop the 5% of observations who chose the decoy. That test is unlikely to matter, but it is still needed.

The analysis we report in the original manuscript is the analysis the reviewer suggests here. That is, we did exclude decoy choices from the analysis. We apologise for the lack of clarity in our original presentation. We have now clarified this in the Exclusion criteria section.

There remains a more substantial problem that could invalidate the general conclusions from the study. Fortunately, there are two simple tests that could determine if the problem exists. The choice design specifies that each respondent sees an A B B' and an A' A B choice set within the same set of tasks. If a respondent

reasonably repeats the first choice between A and B despite the presence of a different undesired decoy, then the average attraction effect will be zero for that respondent, one choice counting for, and the other counting against the attraction effect. To test that possibility, two tests are necessary.

1. Run the analysis only on the first choice made on the A B pair. If recalled repetition is substantial, then the attraction is likely to be positive on the first, but negative on the second choice.

2. Run an analysis on all those who switched A B across the two trials. Contrast the proportion of switches that are in the direction of the attraction effect against those which reverse it.

This is an excellent point and we thank the reviewer for suggesting these analyses. Our results show that participants were indeed unlikely to switch from their first chosen movie, as preferences switched in only 8.5% of trials. To examine this in detail we have performed both of the analyses that the reviewer suggests. The first is to run an analysis only upon trials where subjects encounter a target-competitor pair for the first time. Essentially, mimicking an alternative experimental design where we remove the issue of choice stickiness or consistency by only using each pair once, with one being randomly assigned as the target. This shows no evidence of an attraction effect. The second analysis uses only trials where subjects switch their preferences from the A B A' and A B B' versions. Here again, we see no evidence of a significant attraction effect.

We include the relevant section from the full text below (page 14).

"Our experimental design ensured that for each bespoke A–B movie pair, participants were presented with both A, B, A' and B, A, B' triplets. Faced with two subsequent choices involving two equally highly rated A–B movie pairs and two different, but undesirable decoys, it is possible that the first choice is "sticky", and will be repeated. If this is the case, then we can expect that the target and the competitor will be chosen exactly half of the time (as the target A in A, B, A' is the competitor A in A, B, B'), resulting in a perfectly zero attraction effect.

Indeed, participants were overwhelmingly likely to stick with their first choice, as they only switched between A and B in 8.5% of cases (out of the 990 bespoke A–B pairs where the decoy was not chosen on either of the first or second trial). 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. Thus 57%, 95% CI [46%–68%] of switches were in the direction predicted by the attraction effect, which is not significantly higher than the chance level of 50%, $X^2(1;N=84) = 1.44$, $p = .115$. We have also analysed the first choice participants made for each A–B pair, discarding the second, possibly "sticky" choice. After excluding individual trials where the decoy was chosen, we found no evidence that the proportion of trials where the target was chosen ($M = .49$) was different from .5 on the first choice, $t(134) = .039$, $p = .969$, 95% CI [.46–.52]."

If the attraction effect is still not significant given these two tests, then the paper is an important contribution provided those provided those tests are included. However, if a reliable attraction is found, then there is a valuable paper that identifies the magnitude of the attraction effect and contexts in which it is stronger or weaker. In particular, it may happen that attraction effect becomes diminishes where target-decoy similarity is less or where the respondent are more familiar with the stimuli.

We agree that this is an interesting point. Even though we did not find evidence of a significant attraction effect in either of the two analyses outlined above, we still examine this suggestion. To control for the effect of these potential confounds, we included these variables (familiarity with the movies/target-decoy similarity rating) in our main regression model (Model 2 in Table 1).

It is also important to run the analysis across all the data while adjusting errors to account for within person association. Currently, it is not clear precisely how the authors performed their analysis.

We thank the reviewer for pointing out our lack of clarity in the text. To account for within-subject variability, we used a mixed effects regression framework with subject-specific intercepts. We have now made this clearer.

<u>**Reviewer 2**</u>

 In this paper the authors asked whether using naturalistic stimulus elicits the so-called attraction effect. The design of the study is motivated by mixed views on the robustness of the attraction effect and in particular a claim made by Frederick et al. (2014), according to which the attraction effect occurs only when options are described with numerical attributes. In response to that claim Huber et al. (2014) described another set of criteria that need to be met in order to test for the attraction effect. In the present study, the authors attempt to satisfy both the criteria of Frederick et al. and of Huber et al. and report a zero attraction effect.

Overall the study is both timely and will interest a broad audience. Furthermore, the authors follow an extremely careful procedure in order to construct their choice sets. I believe that this paper will contribute to the ongoing debate regarding the robustness of the attraction effect. However, I can see a few limitations of the current study while I have some requests for further analyses. The latter are needed in order to see if the data consist of a mixture of "repulsion" and "attraction" effects, even within participants, that are determined by different factors (e.g. distance between target and decoy, preference for competitor and target).

Limitations:

1) The authors meticulously elicit similarity and preference ratings for different Netflix moving. These ratings are used in order to construct target-competitor-decoy triplets. Regarding preference ratings, the underlying (reasonable) assumption here is that if two movies receive equal ratings then participants should be indifferent between them. However, I am unclear if this works well in practice. In particular, if choosing between two dissimilar movies of equal ratings would yield 50%-50% choice probabilities. My concern is motivated by the oftentimes discrepancy encountered between judgment and choice experiments but also by the fact that evaluating or comparing dissimilar naturalistic stimuli, such as movies, may engender different cognitive processes. For instance, when asked to rate a movie participants may arrive at the rating by comparing the movie at hand with all other movies within the same genre (e.g. this is a very good action movie). But when asked to compare an action movie and a thriller, their choice could be guided by their overall preferences for one genre over the other (e.g. I strongly prefer a very good thriller over a very good action movie). Ensuring correspondence between equal ratings of dissimilar naturalistic stimuli and choice indifference is necessary in order to claim that the Frederick's criteria are met. This correspondence can only be assessed experimentally. Finally, asides the aforesaid criteria, if preferences elicited from ratings and from choices are decoupled this would undermine the appropriateness of the design in studying the attraction effect in the absence of binary baseline choices between target and competitor.

We thank the reviewer for their very helpful comments and for pointing out this potential issue in the experimental design. To address this, we used average genre ratings as a proxy for overall genre preference. When including this variable in our regression model, we found that participants were marginally more likely to choose the target or competitor with the higher average genre rating. To test whether this pattern influenced the strength of the attraction effect, we repeated the t-test on the subset of trials where participants had similar genre preferences over the target and competitor. We found no evidence for the attraction effect after controlling for these genre effects.

For ease, we have included the additional material from the main text below (page 15).

"Our experimental design relies on the assumption that participants should be indifferent between equally highly rated movies from different genres. However, if the cognitive processes underlying the evaluation and choice stages are different, then discrepancies between ratings and choices might arise. For example, it is possible that the rating reflects preference for the movie within its genre category, but the overall choice between two movies with different genres is driven by overall genre preferences.

To address this concern, we first tested whether overall genre preferences have an influence on choices over and above the information reflected in individual movie ratings, by adding the difference between the target–competitor genre ratings for each participant to our regression models described in Table 1 (Model 1 and 2), assuming that average genre ratings serve as a suitable proxy for overall genre preferences. The results in Table A1 in the Appendix show that overall genre preferences do not change our previous estimate of the attraction effect when added to an intercept-only regression (Model 3). When all of the covariates are included (Model 4) we see no attraction effect, though the effect is less precisely estimated.

Second, to determine whether overall genre preferences influence the strength of the attraction effect, we tested the attraction effect using the subset of trials where overall genre preferences were roughly equal. Using, a one-sample t-test on a subset of trials where absolute difference between the average target and competitor genre ratings was less than 0.25 (about 23% of all trials), we found no evidence that the proportion of trials where the target was chosen (M = .5) was higher than .5, t(98) = .142, p = .556, 95% CI [.46–.55].

Overall, while the results indicate that overall genre preferences slightly influenced choices between the target and competitor, we found no evidence that this had any effect on the strength of the attraction effect."

2) The claims made by Frederick and Huber et al. may not be up to date given more recent experimental results. For instance, Spektor et al. (2018, Psych Science) show that an attraction effect with non-numerical stimuli is obtained only when the alternatives are horizontally aligned (Fig. 5). This speaks to the possibility that when attribute-wise processing is facilitated then the attraction effect ensues and strongly contradicts Fredericks's claim. Interestingly, in all other experiments in Spektor et al., in which the rectangles are not aligned, a repulsion (negative attraction) effect is obtained. It seems, thus, that it matters for the attraction effect whether people engage in attribute- or alternative-wise processing. With naturalistic stimuli, different people may have different strategies which overall gives a zero attraction effect. The paper should reflect the current state-of-the-art (e.g. Spektor et al) beyond the claims made by Frederick et al. and Huber et al.

We thank the reviewer for raising this point, we have now extended the discussion about prior works testing the robustness of the attraction effect in the introduction, and included the suggested highly relevant paper by Spektor et al. along with another recent work by Cataldo and Cohen (page 2).

"The first multiattribute choice experiments demonstrating the attraction effect (e.g., Huber, Payne, & Puto, 1982; Simonson & Tversky, 1992) almost exclusively used stimuli presented as a set of numerical attributes (e.g., cars presented as numerical values for gas mileage and ride quality). Trueblood, Brown, Heathcote, and Busemeyer (2013) have also found evidence for the attraction effect in a perceptual choice experiment, where participants were asked to select the largest from three rectangles with varying widths and heights.

However, recent research suggests that the attraction effect might only occur under very specific conditions. In particular, it had been shown that the effect is much more likely to manifest when an attribute-wise comparison strategy is employed in the choice process, as opposed to an alternative-wise strategy (Noguchi & Stewart, 2014). In addition, the attraction effect seems to be highly dependent on the exact presentation format of the numerical or perceptual choice options (e.g., Spektor, Kellen, & Hotaling, 2018; Cataldo & Cohen, 2019). Since stimulus presentation format fundamentally affects the underlying comparison strategy, a natural concern is then whether this hugely influential decision bias generalises to real-world choice situations, where attributes often cannot be easily visually represented and compared."

Further analyses:

1) The probability of choosing the decoy is low, perhaps too low in comparison to other studies. It is thus an open question whether the decoy was placed way too far from the target, rendering the manipulation ineffective. I recommended plotting the magnitude of the attraction effect against the probability of choosing the decoy (and perhaps the similarity of the two based on the ratings) in order to see if there is any regularity there. The possibility that the decoy was too inferior to generate a preference reversal should be discussed. The mixed effect model ran with similarity ratings shows a lack of effect, but I recommend showing this relationship also descriptively (there might be non-monotonic patterns).

We thank the reviewer for pointing out this potential confound, we agree that it is important to demonstrate that there is no association between the strength of the attraction effect and the preference difference between the target and decoy option, given the work by Spektor et al. on the reverse attraction effect. Furthermore, it is important to allow for the possibility that this relationship could be non-linear. To this end, we calculated the probability of choosing the target from a logistic regression with target-decoy rating difference and similarity rating as explanatory variables. To allow for non-linearity, each potential level of difference was entered as a dummy variable. This analysis found no evidence for an association.

For ease, we include the additional material from the text below (page 16).

"The decoy was chosen very rarely, in less than 5% of trials. Previously, it had been shown that a decoy that is placed too far from the target can result in a reverse attraction effect (repulsion effect; Spektor et al., 2018). On a 1-7 preference rating scale, we allowed for a minimum distance of 3 and a maximum of 6 between the target and decoy. While we have not find any evidence that the target-decoy rating difference influenced the strength of the attraction effect (see Model 2 in Table 1), a

non-linear association between target-decoy preference and the attraction effect might still exist. To examine this possibility whilst controlling for the perceived similarity of the target-decoy pair, we ran a logistic regression model to predict the probability of choosing the target, with target-decoy rating difference and their perceived similarity as explanatory variables. To test for a potential non-monotonic relationship, we estimated a separate coefficient for each level of the explanatory variables. Figure A1 in the Appendix shows the predicted probabilities from this analysis for each combination of target-decoy rating difference and similarity. We found no evidence for the hypothesis that the strength of the attraction effect varies by target-decoy rating difference and similarity rating."

2) How did the preference ratings of the target-competitor (4-4 vs. 7-7) influence the attraction effect? Even if this is part of the logistic model (not sure if it is) I would also recommend to plot this relationship descriptively.

We agree with the reviewer that this is an interesting question, particularly given the role of absolute ratings (over rating differences) in other phenomena such as choice deferral. However, the target-competitor ratings were not included in our pre-registered regression. We now have included additional analyses, with a plot to show the strength of the attraction effect across target-competitor preference ratings.

For ease, we include the additional material from the text below (page 16).

"It is possible that the strength of the attraction effect is influenced by the overall preference for the target and competitor (this is at least 4 and at most 7 in our experiment). To explore this possibility, we calculated the proportion of trials where the target was chosen for each level of target-competitor preference rating. Figure A2 in the Appendix shows these proportions, suggesting that target-competitor preference ratings had no effect on the strength of the attraction effect."

Reviewer 3

Unsurprisingly, I believe the conclusions of this paper, though I have some reservations. Further, since it bears so heavily on this work, I decided to share (almost in full) exchanges I had with reviewers, associate editors, and editors in the long road to getting our paper [Frederick, Lee, & Baskin, 2014] published. (In the first submission, the objections are in red and our responses are in black. In our second submission, the objections are in black and our responses are in red.)

We thank Prof. Frederick for his comments and for sharing the reviews and discussion of Frederick et al. (2014).

p. 2) Regarding our experiments, you write: "participants could even sample the choice options (e.g. squash, mints, popcorn)."   Although the statement is correct, none of the examples are correct, and I'm not sure how you came up with them. (Indeed, I'm not even sure whether squash refers to the vegetable or the sport.)   We had people actually taste kool-aid and actually eat jellybeans and actually feel paper towels and tissue. The remaining stimuli were realistic, but not real.

We thank the reviewer for pointing this out. In part, this was a result of linguistic/cultural differences, e.g. kool-aid is not sold in the UK, but instead we have a similar category of concentrated drink flavouring products called "squash". We have now corrected this section of the paper to accurately (and more globally) represent the stimuli used.

p. 2) I thought the two published objections to our work were weak and I bristle at your genuflection to them here (though I do understand the rhetorical strategy of saying "I've heard you," as well as your desire to differentiate what you did from what we did). I don't think our studies were flawed in the ways these commentators suggested (see my defense in the exchanges below). Accordingly, I object to sentences like "results derived from flawed experiments will not advance our scientific understanding." and attendant suggestions that our work wasn't "rigorous" or "stringent." Similarly, I object to the place in the General Discussion in which you write: "Our experiment is the first investigation to rigorously test the attraction effect with naturalistic stimuli whilst also addressing all of the criticisms raised in connection with Frederick et al.'s experiments." Again, though I acknowledge our stimulus selection/development was more casual than your laudably sophisticated procedures, I think our experiments were mostly just fine and the Huber/Simonson objections mostly just silly. Thus, I don't think it is accurate to say you are the first to do this. Indeed, we didn't even claim to be the first: five other papers before us also used naturalistic stimuli (although we could not replicate those results).

We thank the reviewer for making these points (and for providing us with the revision exchanges), we took his comments on board and removed much of the language in question. However, we stand by the sentence which states *"Our experiment is the first investigation to rigorously test the attraction effect with naturalistic stimuli whilst avoiding the five critical conditions set out by Huber et al. (2014)."* (page 16) We are not aware of any other paper that does address all of the criticisms raised, and therefore this is a factually true statement. But we are not insinuating that the

9

Frederick et al' studies were not rigorous and have adjusted the rest of the text to make this clearer!

p. 4) It is awkward to write "movies that are part of the same sequel" (a sequel is single movie). You might say "part of a series" p. 4) You write "it is unclear how Frederick et al. decided which movie should be the target." Let me clarify, then. The customary nomenclature in tests of AD involve the target and competitor as the two "legitimate" options that one might sensibly choose and the decoy as the option which most closely resembles, but is inferior to, the target. Hence the identity of the decoy essentially defines the identity of the target. The assignation is quite clear for nearly all our stimuli. Further, while we did assume that the sequel functioned as the decoy in our stimuli, it is important to note that the data support the assumption (it is rarely chosen).

We thank the reviewer for the clarification. We have now reworded this paragraph to highlight the differences between our respective experimental designs (page 5).

"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. Frederick et al. (2014) have also used movie stimuli in two of their experiments: they chose pairs of movies that are part of the same series or are starring the same actor (but have distinctly different genres) to create target-decoy pairs. In these experiments, the identity role of each of the three movies (target, competitor, decoy) was always the same for all participants, and based upon population average ratings rather than individual ratings."

p. 6) It took me a very long time to comprehend what you are trying to do in stimuli creation. The description in the Experimental procedure question seems much clearer than the description that precedes it.   By the end, I liked what you did, but I think/hope it can be expressed more simply.

We thank the reviewer for pointing this out, indeed, our quadruplet construction process is rather complex. We have now added a flowchart (Figure 2), to provide a summary of the steps involved in this process. We hope this provides a useful overview of the process.

p. 8) "We did not collect any data about the demographics of our sample, as we did not expect it to affect our results." Can I just say how refreshing it is to read this?! I commend you on having the common sense and courage to just say this. Good for you. I hope it catches on.

p. 12) Can you clarify what you mean when you write that "we hypothesize that spatial separability of the attribute dimensions might be a key role in the comparison process." Also, although you've obviously read our paper, it seems like you haven't when you write: "Future research could test this hypothesis this [sic] by exploring how the strength of the attraction effect varies with different representations of the same choice options (e.g. numerical, visual)." This is literally exactly what we did in in our paper in studies 2a through 2c and studies 3a through 3c.

We thank the reviewer for pointing this out, we have now re-worded the paragraph to clarify our point (page 18). There is considerable evidence (e.g., Cataldo & Cohen, 2019; Noguchi & Stewart, 2014) showing that attribute-wise processing of the choice options is key to the attraction effect. We speculate that different representations of the same stimuli (more specifically, of its attribute dimensions) can give rise to different comparison processes, influencing the strength of the attraction effect. While many studies have demonstrated that the attraction effect is largely limited to choice settings with numerical attributes, future studies investigating how the comparison process changes with different representations of the same underlying stimuli could inform us about *why* the attraction effect is only present under very specific circumstances.

"In conclusion, our results are in line with that of Frederick et al. (2014), and provide strong evidence that the attraction effect does not extend to choice between naturalistic options. While we did not aim to investigate the exact reason behind why the attraction effect is robust in choices involving numerical attribute dimensions but is absent from choices involving naturalistic options, we hypothesize that the separability of the attribute dimensions is an important factor. For example, attribute information is spatially separate in a numerical matrix form presentation, but attribute information occurs in the same spatial location for movie thumbnails. Given the evidence that attribute-wise comparison strategies are key to the attraction effect in numerical or perceptual choices, such comparison strategies are less likely to occur with complex, naturalistic objects where the attributes are not spatially separate. Future research could test this hypothesis by exploring how the comparison process and the strength of the attraction effect varies with different representations of the same choice options (separate attributes versus naturalistic representation), extending Frederick et al.'s investigation of numerical versus visual presentation of the same stimuli. Results from such experiments could provide us with important insights about the boundary conditions of the attraction effect, and the cognitive process underlying this choice bias."

In general, despite my predisposition to like this paper, and my admiration of some of the design features, I confess I have a hard time following exactly what was done. Maybe that requires a diagram or something. If I understand your procedure correctly (and I'm not sure I do), respondents first encountered the {A,B,b} triplet and later encountered the {A,B,a} triplet and you examined whether the nature of the decoy (a or b) affected the choice between the core options (A or B). Even with the many interceding "filler" choices, this seems like a weak test of the potential power of the decoy option, in much the same way that it would be a weird test to ask respondents:

Which do you prefer: A or B? and, then immediately after, Which do you prefer: A or B or b? That said, I think there is an opportunity here for both a within-subjects analysis (seeing whether a person changes their choices with different decoys) and a between-subjects analysis (seeing whether people, on average, make different choices with different decoys). Assuming item order was randomized, I think both types of analysis are possible, and I would like to see both.

We thank the reviewer for his comments. As described above, we have now included a "further analysis" section, where we include tests of the attraction effect using only the first or second trial corresponding to each quadruplet (half the overall trials). In addition, our regression analyses are all within-subjects analyses including subject-specific intercepts.