nature research

Peer Review Information

Journal: Nature Human Behaviour
Manuscript Title: Revealing the multidimensional mental representations of natural objects underlying human similarity judgments
Corresponding author name(s): Martin Hebart

Editorial Notes:

Reviewer Comments & Decisions:

Decision Letter, initial version:

4th May 2020

*Please ensure you delete the link to your author homepage in this e-mail if you wish to forward it to your co-authors.

Dear Martin,

Thank you once again for submitting your revised manuscript, entitled "Revealing the multidimensional mental representations of natural objects underlying human similarity judgments," and for your patience during the re-review process.

Your manuscript has now been evaluated by our referees, and in the light of their advice I am delighted to say that we can in principle offer to publish it. First, however, we would like you to revise your paper to address the points made by the reviewers, and to ensure that it complies with our Guide to Authors at http://www.nature.com/nathumbehav/info/gta.

Nature Human Behaviour offers a transparent peer review option for new original research manuscripts submitted from 1st December 2019. We encourage increased transparency in peer review by publishing the reviewer comments, author rebuttal letters and editorial decision letters if the authors agree. Such peer review material is made available as a supplementary peer review file. Please state in the cover letter 'I wish to participate in transparent peer review' if you want to opt in, or 'I do not wish to participate in transparent peer review' if you don't. Failure to state your preference will result in delays in accepting your manuscript for publication.
Please note: we allow redactions to authors' rebuttal and reviewer comments in the interest of confidentiality. If you are concerned about the release of confidential data, please let us know specifically what information you would like to have removed. Please note that we cannot incorporate

redactions for any other reasons. Reviewer names will be published in the peer review files if the reviewer signed the comments to authors, or if reviewers explicitly agree to release their name. For more information, please refer to our FAQ page.

We ask you to revise your manuscript to improve the clarity of descriptions and include the necessary explanations of or justification for your approach, following the constructive comments made by Reviewer #1 and Reviewer #2 in that regard. In particular, I highlight the request to respond to Reviewer #1's second point, and Reviewer #3's third request.

One of the main reasons for delays in formal acceptance is failure to fully comply with editorial policies and formatting requirements. To assist you with finalizing your manuscript for publication, I attach a checklist that lists all of our editorial policies and formatting requirements. I also attach a template document, which exemplifies our policies and formatting requirements.

Please attend to *every item* in the checklist and upload a copy of the completed checklist with your submission. I have highlighted in the checklist items that require your attention. I also mention here a few points that are frequently missed and can cause delays:

1) Ensure that all corresponding authors have linked their ORCID to their account on our online manuscript handling system. This is very frequently missed and invariably causes delays in formal acceptance.

2) Ensure that you provide all of the materials requested in the attached checklist and below with your final submission. Please note that the Licence to Publish needs to be hand-signed.

3) Please reconsider use of Supplementary Material, instead opting to make use of Extended Data, which will considerably improve the accessibility and transparency of your report. On the attached checklist, I have highlighted some specific recommendations.

We hope to hear from you within 10 days; please let us know if the revision process is likely to take longer.

To submit your revised manuscript, you will need to provide the following:
• Cover letter
• Point-by-point response to the reviewers (if applicable)
• Manuscript text (not including the figures) in .docx or .tex format
• Individual figure files (one figure per file)
• Extended Data & Supplementary Information, as instructed
• Reporting summary
• Editorial policy checklist
• License to publish
• Third-party rights table (if applicable)
• Suggestions for cover illustrations (if desired)

Consortia authorship:
For papers containing one or more consortia, all members of the consortium who contributed to the

paper must be listed in the paper (i.e., print/online PDF). If necessary, individual authors can be listed in both the main author list and as a member of a consortium listed at the end of the paper. When submitting your revised manuscript via the online submission system, the consortium name should be entered as an author, together with the contact details of a nominated consortium representative. See https://www.nature.com/authors/policies/authorship.html for our authorship policy and https://www.nature.com/documents/nr-consortia-formatting.pdf for further consortia formatting guidelines, which should be adhered to prior to acceptance.

Please use the following link for uploading these materials:
[REDACTED]

If you have any further questions, please feel free to contact me.

With best regards,

Anne-Marike Schiffer
Editor
Nature Human Behaviour

Reviewer #1:
Remarks to the Author:
The manuscript by Hebart et al. presents a computational model of mental representations of objects based on a large-scale assessment of human similarity judgments for natural object images (taken from the authors' previously published THINGS image database). The topic of how the human mind represents object is a highly interesting and much-debated one, with prior research generally having produced evidence for only few interpretable dimensions (e.g., animacy or size), and beyond that a general hot mess.

A noteworthy strength of the authors' approach is its reliance on a large online sample (consisting of 1.46 millions responses from around 5,000 mTurk participants, validated with 20 in-lab subjects) together with its clever experimental design and computational sophistication. Specifically, the authors employed a clever triplet odd-one-out task that enabled them to get similarity ratings for 1,854 objects without a priori constraining these ratings to particular visual or conceptual dimensions. The similarity ratings were then used to train a shallow neural network to map individual objects to a 90-dimensional output vector, with sparseness and connectivity constraints. After training, the authors used a cutoff to eliminate features with low weights, resulting in 49 dimensions along which participants appeared to compare objects. Rather strikingly (and in noteworthy contrast to other recent approaches, e.g., of Huth et al. Nature, 2016, who have claimed representational schemes consisting of dimensions which in general turned out to be rather hard to interpret), most of these dimensions could easily be labeled by human participants, with high inter-subject agreement. Computational controls showed that most of these dimensions (34/49) were robust under varying starting conditions (in this context, it would be interesting if the authors could discuss a bit more which dimensions showed greater variability – do these more variable dimensions share particular properties? Is the variability in the computational analyses in any way related to variability in subjects' labeling choices?).

The authors then report performance of the model on an independent test set of 48 objects whose similarity matrix was completely sampled in another mTurk experiment. Quite excitingly, the model was able to predict behavioral similarities with an accuracy close to the noise ceiling – a feat that, to my knowledge, has never been accomplished by prior representational schemes.

Further adding to the strengths of the study and the authors' model of object similarity, Hebart et al. show that the dimensions found through their analyses could even be used to _generate_ arbitrary pairwise similarity ratings by having subjects rate _individual_ objects along the dimensions of the embedding.

In summary, the study by Hebart et al. represents a breakthrough in our study of how humans represent objects by presenting a multi-dimensional model of unprecedented explanatory power and interpretability. The research directly suggests a host of intriguing follow up questions (e.g., can neural correlates of the different representational dimensions be found – which might then be differentially affected by specific brain lesions? How universal are the representational dimensions across different cultures? Age groups? SES? etc. etc.), and should be of substantial interest to a number of readers from different fields.

While already a very strong manuscript, in terms of suggestions for improvements in a revision it would be good if the authors could (in addition to discussing questions already raised above) provide more detail on their mTurk sample. Where were the workers located? What was their age & gender distribution? Is there are any information on their race/ethnicity, education and/or SES, all of which might modulate their mental representations? Could all participants be assumed to be familiar with all the objects tested? Similar questions apply to the in-lab sample, where only the gender distribution is given in the manuscript, but no information even about participant age. Was the in-lab sample matched to the mTurk sample in any way? Finally, regarding the computational analyses, what was the justification for the initial choice of 90 dimensions? Did the authors explore larger and smaller number of starting dimensions (in particular less than 49)? What were the results?

Minor comment:
The bibliography had a number of issues (e.g., ref. 30 was incomplete, ref. 46 missing page numbers; in general, style was inconsistent: some papers are cited without page numbers, some with a starting page number, some with first and last page number)

Reviewer #2:
Remarks to the Author:
General comments

In this paper, the authors aimed to identify core dimensions according to which we are able to carry out similarity judgements of objects. In contrast to previous approaches that were agnostic to the relevant properties or dimensions, the authors used a data-driven computational model of similarity judgments for pictures of 1,854 objects. The authors obtained 49 object dimensions, such as colourful, disc-shaped, or food-related. The authors demonstrated that these object dimensions were interpretable and reproducible, and that they predicted behavioural performance in terms of categorization behaviour and typicality ratings.

This study has been carefully carried out and addresses a very relevant topic that clearly emerges from the corresponding literature on object representations while overcoming several limitations of previous studies. In the view of this reviewer, the data-driven approach used by the authors provides an important step forward in understanding the properties according to which we categorize objects, and provides the methodological and conceptual basis for asking similar questions in other domains (e.g. words, faces, or places). My comments mostly refer to methodological details that I would like the authors to clarify/ describe in more detail for ease of understanding.

(1) Page 2, last paragraph: I'm not entirely clear on how the authors got from 1.06 billion possible combinations to 1.46 million unique responses. Likewise, what exactly was the relationship between the fully sampled matrix of 48 objects (done by 121 workers) and the data used for training and testing the computational model (done by 5,301 workers)? It would be helpful to expand on this aspect in the corresponding methods section.

(2) Page 4, first paragraph: Unless I misunderstood, rows correspond to object vectors and columns to dimensions, not the other way around.

(3) Page 4: I'm not sure about the assumption of dimensions being continuous. As an example, wouldn't one consider a dimension such as animacy to be binary rather than continuous? If so, how would this affect the interpretation of the results?

(4) Page 4, last paragraph: The authors may want to refer to the Methods section for further details.

(5) Page 5, first paragraph: The authors may want to make explicit that these 1,000 triplets were chosen from the same original database.

(6) Page 9, first paragraph: For readers less familiar with t-distributed stochastic neighbourhood embedding, the authors should provide some more details regarding how they projected the 49-dimensional similarity embedding to 2 dimensions.

(7) Page 10, second paragraph: The authors may want to provide more details on the procedure used to predict category membership for each of the 1,112 objects of the categories.

(8) Figure 8, panel a: Were all of these example images shown for one single dimension, and if so, what was the name of that dimension? It would be helpful to clarify this, even if the label wasn't provided to the participants.

(9) Related to the previous point, it would be helpful to know which exact instruction was provided to the participants. In the example provided in Figure 8a, I wouldn't be quite sure where on the scale to place the image of the flamingo – 'not at all', because there is another bird? But then, what does the bird have in common with toast, spring onions, nuts and coffee?

(10) Page 18, second paragraph: Please provide details on how you chose the 48 objects. As an example, how did you arrive at the word vectors that were used for spectral clustering? According to which criteria was one object per cluster selected?

(11) A conceptually related study the authors might want to discuss in the context of the current

study is the paper by Watson and Buxbaum (2014, JEP: HPP) that used a data-driven approach to reveal the key dimension underlying the organization of tool-use actions.

Author Rebuttal to Initial comments

Response to Reviewers for Manuscript "Revealing the multidimensional mental representations of natural objects underlying human similarity judgments"

(please note that page numbers for the marked changes below refer to the document including highlights and Figures, not the final submission without highlights or Figures)

Response to Reviewer #1:

We would like to thank Reviewer #1 for their positive assessment of our work and their helpful suggestions for improvements and clarifications.

R#1: Computational controls showed that most of these dimensions (34/49) were robust under varying starting conditions (in this context, it would be interesting if the authors could discuss a bit more which dimensions showed greater variability – do these more variable dimensions share particular properties?

In response to the reviewer's request, we sorted the dimensions based on their variability and inspected the labels. The 15 dimensions with the lowest consistency are (starting with the least consistent): "cylindrical / conical", "handicraft-related", "container-related / hollow", "has beams / support", "construction-related", "has grating", "thin / flat", "black / noble", "feminine (stereotypically)", "repetitiveness", "bathroom-related", "arms/legs/skin-related", "medicine-related", "long-thin", "shiny / transparent". These dimensions do not seem to be dominated by particular types of dimensions such as perceptual or conceptual. However, these dimensions tended to be those with a lower overall weight summed across all objects (this can also be inferred from Supplemental Figure 1). As a reminder, we had sorted the 49 object dimensions based on their overall weight across all objects, in descending order. The rank of the dimension reliability exhibited a strong positive correlation with the rank of the dimension (Spearman's ρ: 0.75, $p < 0.001$, randomization test). This result makes sense since changes in these dimensions would have a smaller impact on the overall predictions. Were there more objects that shared these dimensions, we would expect them to be more stable.

Page 5: “There was a strong correlation between the ranks of the dimensions and the dimension reproducibility (Spearman’s ρ: 0.75, $p < 0.001$, randomization test), indicating that reproducibility of individual dimensions was driven mostly by their overall importance in the model.”

R#1: Is the variability in the computational analyses in any way related to variability in subjects’ labeling choices?

This is an interesting question. Unfortunately, it is difficult to quantify the variability in subjects’ labeling choices, since they often used very similar, but slightly different words for the same meaning. We considered using word embeddings or sentence embeddings as a proxy, but this approach makes additional assumptions that need not hold in practice. In addition, the $n = 20$ for the in-lab participants is a limiting factor for quantifying this variability across participants. At the same time, the effect described in the response to the previous question (overall relevance of dimensions) seems to capture much of this variability in computational analyses already.

R*#1: It would be good if the authors could [...] provide more detail on their mTurk sample. Where were the workers located? What was their age & gender distribution? Is there any information on their race/ethnicity, education and/or SES, all of which might modulate their mental representations?*

We thank the reviewer for these suggestions. Individual variability based on age, gender, geography, or other sociodemographic factors is a highly interesting question for future studies, which we now discuss more explicitly in the outlook section of the manuscript. In addition, the manuscript now includes gender in the results section and states that the location of the workers was restricted to the USA. Self-identified race/ethnicity information was collected on a voluntary basis as a requirement by NIH, but was not collected for the purpose of data analysis. Unfortunately, when conceiving the study, we had not planned to examine any potential interindividual variability. For future datasets, we will be sure to include age as an additional variable and have added the following when we discuss future directions.

Page 18: “To what degree are the dimensions shared between different individuals, and how are they affected by gender, age, culture, education, other sociodemographic factors, and individual familiarity with the objects?”

R#1: Could all participants be assumed to be familiar with all the objects tested?

We chose the objects based on the 1,854 objects in the THINGS database (Hebart et al., 2019, PLoS ONE). All of the object images had been named by test participants, and only images that were named consistently were included (as defined in the article describing the database). We now explicitly mention this in the manuscript. However, there is still a possibility that workers in the current study had variability in their familiarity with the objects. In the instructions, participants were told that if they did not recognize the object, they should base their choice on their best guess. We now include more details on the instructions to participants. Finally, we mention familiarity with the object as a potential factor worth addressing in future studies when looking at individual differences (see previous response for changes in text).

Page 19: “Importantly, the validation task of the THINGS database demonstrated that the objects in the 1,854 images were generally nameable, i.e. it can be assumed that most participants were sufficiently familiar with the objects to be able to name them.”

Page 19: “In addition, participants were instructed that in case they did not recognize the object, they should base their judgment on their best guess of what the object could be.”

R#1: Similar questions apply to the in-lab sample, where only the gender distribution is given in the manuscript, but no information even about participant age. Was the in-lab sample matched to the mTurk sample in any way?

Apologies, we noticed these data were missing before submission but due to COVID-19 we temporarily had difficulty accessing this information. The age distribution is now included in the manuscript. The in-lab sample was not matched to the mTurk sample.

R#1: Regarding the computational analyses, what was the justification for the initial choice of 90 dimensions? Did the authors explore larger and smaller number of starting dimensions (in particular less than 49)? What were the results?

As mentioned in the results section of the manuscript, the choice of 90 dimensions was chosen based on the assumption that a smaller number of dimensions would be sufficient. Had we found that none of the dimensions are 0, we would have re-run the model with more dimensions.

What is the effect of initializing with more dimensions? We once ran the model with 200 dimensions, with no discernible effect on the final solution in terms of number of dimensions or

model fit. The benefit of the sparsity constraint (L1-norm) is that it can deal with too many dimensions quite well. L1-penalized models are known to be robust to up to exp(n) dimensions. This means the model could easily deal with 1000 dimensions and would very likely yield a highly similar solution. We now discuss this more explicitly.

What is the effect of initializing with fewer dimensions? We expect the resulting dimensions to be less interpretable. We once ran a model without the sparsity constraint. This led to a model that still performed very well in predicting individual choices, but to non-interpretable dimensions. Choosing a model with fewer dimensions would mean we would reduce the L1-penalty, i.e. dimensions would turn out to be less sparse and more similar to the non-sparse model we ran. For that reason, we expect the dimensions to be less interpretable.

Page 19: “Note that initializing the model with 200 dimensions led to very similar model performance and final number of dimensions (results not shown).”

R#1: The bibliography had a number of issues (e.g., ref. 30 was incomplete, ref. 46 missing page numbers; in general, style was inconsistent: some papers are cited without page numbers, some with a starting page number, some with first and last page number)

We thank the reviewer for noticing these issues, which we corrected in the revised manuscript.

Response to Reviewer #2:

We thank Reviewer #2 for their positive evaluation and for asking for methodological clarification, which we are happy to provide.

R#2 (1) Page 2, last paragraph: I'm not entirely clear on how the authors got from 1.06 billion possible combinations to 1.46 million unique responses. Likewise, what exactly was the relationship between the fully sampled matrix of 48 objects (done by 121 workers) and the data used for training and testing the computational model (done by 5,301 workers)? It would be helpful to expand on this aspect in the corresponding methods section.

We would like to thank the reviewer for pointing out the need to clarify this. The ~1.46 million unique responses were obtained with triplets drawn at random from the ~1.06 billion possibilities. The number of responses collected was determined solely by the logistics of acquisition. The practical questions we posed ourselves were (1) whether this sufficed for producing an embedding that would work in predicting individual behavioral choices for unseen triplets and (2) how well this would work for approximating behavioral similarity, here defined by the probability of choosing two objects *i* and *j* together in the triplet task across all possible third objects *k*.

To address the first question, we split our measured triplet data into training and test data and measured the predictive performance of individual triplet choices. To address the second question, we considered 48 objects and *every possible* triplet assembled from them (48!/(45!3!) = 17,296) and collected a separate dataset with multiple ratings for each of these. This, in turn, gave us a fully sampled similarity matrix determined by behavior. Given those triplets, we used the embedding to predict behavior in them, and derived a predicted similarity matrix. The degree to which the 48 object similarity matrix could be approximated provided us with an indication that the embedding captured sufficient information about similarity. We now improve the description of these aspects in the manuscript.

Caption of Figure 1: "Since only a subset of all possible triplets had been sampled (0.14 % of 1.06 billion possible combinations), this model additionally served to complete the sparsely sampled similarity matrix."

Caption of Figure 2: "Predictiveness of the computational model for single trial behavioral judgments and similarity."

Page 6: “Since we had sampled only a fraction of the 1,854 × 1,854 similarity matrix, the test data were insufficient for addressing how well the model could predict behaviorally measured similarity.”

Page 20: “For the fully-sampled similarity matrix of 48 objects used for testing the performance of the model at predicting object similarity (Fig. 2), we created a different similarity matrix that was constrained only by this subset of 48 objects.”
R#2 (2) Page 4, first paragraph: Unless I misunderstood, rows correspond to object vectors and columns to dimensions, not the other way around.

We thank the reviewer for spotting this mistake, which we fixed.

R#2 (3) Page 4: I’m not sure about the assumption of dimensions being continuous. As an example, wouldn’t one consider a dimension such as animacy to be binary rather than continuous? If so, how would this affect the interpretation of the results?

Indeed, if dimensions were assumed to be binary features as is the case for well-described object property norms (e.g. “has legs”, “does fly”, “is animate”), then it would make sense for dimensions to be modeled as binary. In contrast, the dimensions here are supposed to reflect the graded nature of these object properties, rather than our categorical judgments / semantic knowledge of them. Property judgments are easy for many objects, but are more difficult for others. For example, do ants fly? They do, but it is not their most common or typical property. Do birds fly? This is a lot easier to answer. And while we can categorize objects as being animate or inanimate, many people would likely agree in a forced judgment that a lion is more animate than a star fish (see Carlson et al., 2013, Journal of Cognitive Neuroscience, which is now included as a reference in the manuscript). This reasoning was underlying our choice for choosing dimensions to be continuous and was confirmed in our typicality experiment. We now expand on the explanation of this reasoning in the discussion section.

Page 17: “Analyses of category-related typicality judgments demonstrate that the continuous nature of the dimensions is informative as to the degree to which these dimensions are expressed in objects, demonstrating that continuous dimensions allow us to generalize beyond binary categorical assignment of semantic attributes (e.g. “is animate”).”

R#2 (4) Page 4, last paragraph: The authors may want to refer to the Methods section for further details.

This has now been included in the revised manuscript.

R#2 (5) Page 5, first paragraph: The authors may want to make explicit that these 1,000 triplets were chosen from the same original database.

We now make explicit that these 1,000 triplets were chosen from the set of those for the 1,854 objects.

R#2 (6) Page 9, first paragraph: For readers less familiar with t-distributed stochastic neighbourhood embedding, the authors should provide some more details regarding how they projected the 49-dimensional similarity embedding to 2 dimensions.

We now provide more details in the main text.

Page 10: " First, we projected the 49-dimensional similarity embedding to 2 dimensions using *t*-distributed stochastic neighborhood embedding (*t*-SNE, dual perplexity: 5 and 30), initialized using metric multidimensional scaling. This approach has been shown to preserve the global similarity structure while providing a higher degree of interpretability at the local similarity level than multidimensional scaling alone."

R#2 (7) Page 10, second paragraph: The authors may want to provide more details on the procedure used to predict category membership for each of the 1,112 objects of the categories.

We now provide more details in the Methods section.

Pages 20/21: "These categories comprised 1,112 objects. Classification was carried out using leave-one-object-out cross-validation. For training, Centroids for all 18 categories were computed by averaging the 49-dimensional vectors of all objects in each category, excluding the left-out object. The membership of this remaining object was then predicted by the smallest Euclidean distance to each centroid. This procedure was repeated for all 1,112 objects, and prediction accuracy was averaged. For the corresponding analysis with a semantic embedding, we used a publicly-available 300-dimensional sense embedding."

R#2 (8) Figure 8, panel a: Were all of these example images shown for one single dimension, and if so, what was the name of that dimension? It would be helpful to clarify this, even if the label wasn't provided to the participants.

Most of these example images were shown for this dimension (due to copyright issues, we replaced a small number of images by very similar images). For each dimension, different images were shown. We now clarify the name of this dimension, that the specific images varied between dimensions, and that all 20 participants rated all 20 images.

R#2 (9) Related to the previous point, it would be helpful to know which exact instruction was provided to the participants. In the example provided in Figure 8a, I wouldn't be quite sure where on the scale to place the image of the flamingo – 'not at all', because there is another bird? But then, what does the bird have in common with toast, spring onions, nuts and coffee?

We now expand on the description of the instructions in the Methods section ("Dimension naming task" and "Object dimension rating task"). To answer the reviewer's question, what these objects have in common is the absence of the dimension. In other words, low values in a dimension do not carry any meaning (although for the shown dimension "artificial / hard", natural objects tended to fall at the bottom of this dimension).

R#2 (10) Page 18, second paragraph: Please provide details on how you chose the 48 objects. As an example, how did you arrive at the word vectors that were used for spectral clustering? According to which criteria was one object per cluster selected?

We now provide more details.

Page 19: "To yield a diverse set of objects for the fully sampled reference dataset, the 48 objects were chosen by carrying out spectral clustering on publicly-available 300-dimensional sense vectors of all 1,854 objects with 48 clusters and by choosing one object per cluster randomly."

R#2 (11) A conceptually related study the authors might want to discuss in the context of the current study is the paper by Watson and Buxbaum (2014, JEP: HPP) that used a data-driven approach to reveal the key dimension underlying the organization of tool-use actions.

We thank the authors for bringing up this relevant work. While we considered to include it, we noticed that we barely cited any previous experimental work using data-driven approaches, and given the limited space focused on larger reviews. Hence, we believe that it would be rather selective to discuss this work and not other work in similar domains such as faces. We hope the reviewer is understanding of our decision.

Decision Letter, first revision:

12th August 2020

Dear Martin,

Thank you once again for your manuscript, entitled "Revealing the multidimensional mental representations of natural objects underlying human similarity judgments".

I'm pleased to see that all documentation is now in order, so that this does no longer stand in the way of publication. However, editorially, we do require a significant change to the way you responded to an editorial concern before we can make a final decision on publication.

In my last decision letter, I requested that all instances of "results not shown" were removed from the manuscript, instead presenting all relevant results in the manuscript or SI. Extended Data Figure 1 presents an appropriate response to the referee concern regarding the number of dimensions at initialization, for which results were previously not included.

In another instance, addressing Reviewer #2's point 3, you now cite reference 33 (Zheng, C. Y., Pereira, F., Baker, C. I. & Hebart, M. N. (2019) Revealing interpretable object representations from human behavior. https://arxiv.org/abs/1901.02915), a preprint that seems to represent effectively an earlier version reporting a subset of the present work.

Although references to preprints are generally permitted, referring to a preprint of the same work is not a suitable solution. In the preprint, you argue why a dimensional approach is more suitable than a binary approach, and demonstrate that individual dimensions represent complex combinations of the information in the binary features in Devereux et al. (2014). In your final revision, we request that you instead incorporate any empirical evidence in the SI, and include the arguments supporting your case in the manuscript.

Once this final request has been addressed, we will be able to accept your manuscript for publication (without further delay).

Please use the link below to submit the finalized version of manuscript and related files:

[REDACTED]

Note: This URL links to your confidential home page and associated information

about manuscripts you may have submitted, or that you are reviewing for us. If you wish to forward this email to co-authors, please delete the link to your homepage.

Many thanks in advance and I look forward to receiving the finalized manuscript.

Best wishes,
Marike

Marike Schiffer, PhD
Senior Editor
Nature Human Behaviour

Author Rebuttal, first revision:

Second Rebuttal Letter for Manuscript "Revealing the multidimensional mental representations of natural objects underlying human similarity judgments"

We would like to thank the Senior Editor (Anne-Marike Schiffer) for her continued interest and support of the manuscript. We have now addressed the outstanding issues raised by her, as well as additional issues that we became aware of in the process. We will detail these issues and the changes to the manuscript below.

The editor expressed concern regarding our response to Reviewer #2's third comment, specifically the citation used in our response, which referred to our own work (Zheng et al., 2019) available as a preprint and presented at a conference:

"*It may be possible to generate these binary properties from the continuous dimensions in our model*[33]*, which would demonstrate that the implicit judgments in the odd-one-out task capture much of the explicit semantic knowledge of objects, but a general test of this idea would require the creation of feature production norms for the 1,854 objects used in the creation of the embedding.*"

The reason we chose this reference was that our previous work included similar deliberations. However, since there is a lot of overlap between the content of this preprint and the present work and since the presented evidence in the preprint was not very strong, in response to the editor's concern, we now remove this citation. Note that the removal of this citation neither affects the content of the sentence nor should it affect the response to the reviewer's concerns.

In addition, we removed another reference to this preprint, where we cited evidence that a model with Euclidean distance rather than the dot product exhibited similar performance. We now include these results in the manuscript text:

"*The dot product was chosen as a basis for proximity for computational reasons, but using Euclidean distance led to similar performance (prediction accuracy of test set odd-one-out choices: 64.69%, dimensions: 57).*"

Finally, there was a remaining mention of "results not shown" in the manuscript that we missed in our previous revisions. We now replaced this by a reference to a new Extended Data Figure 4.

"*While we demonstrated that most dimensions were highly reproducible across different random initializations of the model, using a smaller subset of the data for building the embedding revealed a smaller number of dimensions (Extended Data Figure 4), indicating that the dimensionality of the embedding is a function on the amount of data used.*"

Final Decision Letter:

Dear Martin,

We are pleased to inform you that your Article "Revealing the multidimensional mental representations of natural objects underlying human similarity judgments", has now been accepted for publication in Nature Human Behaviour.

Before your manuscript is typeset, we will edit the text to ensure it is intelligible to our wide readership and conforms to house style. We look particularly carefully at the titles of all papers to ensure that they are relatively brief and understandable.

The subeditor may send you the edited text for your approval. Once your manuscript is typeset you will receive a link to your electronic proof via email within 20 working days, with a request to make any corrections within 48 hours. If you have queries at any point during the production process then please contact the production team at rjsproduction@springernature.com. Once your paper has been scheduled for online publication, the Nature press office will be in touch to confirm the details.

Acceptance of your manuscript is conditional on all authors' agreement with our publication policies (see http://www.nature.com/nathumbehav/info/gta). In particular your manuscript must not be published elsewhere and there must be no announcement of the work to any media outlet until the publication date (the day on which it is uploaded onto our web site).

The Author's Accepted Manuscript (the accepted version of the manuscript as submitted by the author) may only be posted 6 months after the paper is published, consistent with our <a

href="http://www.nature.com/authors/policies/license.html">self-archiving embargo. Please note that the Author's Accepted Manuscript may not be released under a Creative Commons license. For Nature Research Terms of Reuse of archived manuscripts please see: http://www.nature.com/authors/policies/license.html#terms
If you have posted a preprint on any preprint server, please ensure that the preprint details are updated with a publication reference, including the DOI and a URL to the published version of the article on the journal website.

An online order form for reprints of your paper is available at https://www.nature.com/reprints/author-reprints.html. All co-authors, authors' institutions and authors' funding agencies can order reprints using the form appropriate to their geographical region.

We welcome the submission of potential cover material (including a short caption of around 40 words) related to your manuscript; suggestions should be sent to Nature Human Behaviour as electronic files (the image should be 300 dpi at 210 x 297 mm in either TIFF or JPEG format). Please note that such pictures should be selected more for their aesthetic appeal than for their scientific content, and that colour images work better than black and white or grayscale images. Please do not try to design a cover with the Nature Human Behaviour logo etc., and please do not submit composites of images related to your work. I am sure you will understand that we cannot make any promise as to whether any of your suggestions might be selected for the cover of the journal.

You can now use a single sign-on for all your accounts, view the status of all your manuscript submissions and reviews, access usage statistics for your published articles and download a record of your refereeing activity for the Nature journals.

To assist our authors in disseminating their research to the broader community, our SharedIt initiative provides you with a unique shareable link that will allow anyone (with or without a subscription) to read the published article. Recipients of the link with a subscription will also be able to download and print the PDF.

As soon as your article is published, you will receive an automated email with your shareable link.

We look forward to publishing your paper.

best
Marike

Marike Schiffer, PhD
Senior Editor
Nature Human Behaviour

P.S. Click on the following link if you would like to recommend Nature Human Behaviour to your librarian http://www.nature.com/subscriptions/recommend.html#forms