metareviewer review (reviewer 4)  
score 3/5  
  
  Type of Submission  
  
    Cognitive Science  
  
  The Review  
  
    The paper expands previous work on alternating category structure using a single-  
    dimension to a two-dimensional space. The paper also presents a neural net model  
    that is generally able to capture the results. The three reviewers are in  
    agreement that the work presented is good, interesting, and relevant to the CogSci  
    community and I share their enthusiasm. However, I do believe the work is rather  
    incremental and perhaps more appropriate for a poster. The reviewers also point  
    out a series of issues that I believe are worth addressing and taking into account  
    as the authors continue with this line of research.  
  
----------------------------------------------------------------  
  
reviewer review (reviewer 1)  
score 3/5  
  
  Type of Submission  
  
    Cognitive Science, Psychology  
  
  The Review  
  
    This paper investigates whether participants can learn categories that alternate  
    over a stimulus space. I thought this paper reported interesting and novel  
    results, and addressed a theoretical question of significance (whether and how  
    people learn relational categories). The experiments were mostly well-designed and  
    the analyses were appropriate and thorough. Many of my comments are points of  
    clarification/presentation.  
  
    I think the 2D category design used by the authors is a significant extension on  
    Kurtz and Wetzel (2021). Although the other dimension was only partially  
    predictive, the 1D rule is much simpler to learn. Any transfer based on  
    alternation is more surprising in the presence of this simpler 1D rule (but lack  
    of counterbalancing is a limitation, see below). I think the novelty of this  
    design choice could be emphasised, and given more discussion. For example, is  
    there anything in the data that tells us whether the participants who alternated  
    learned anything about the other dimension? Perhaps alternations are harder to  
    learn but dominate behavior once learned. Do the authors think that they would  
    have obtained the same results if the dimensions were counterbalanced (and why  
    weren’t they)?  
  
    On the top of page 3, different testing thirds are described, but I wasn’t sure  
    how the testing space was divided into these thirds. Perhaps Figure 3 would be  
    marked accordingly. The captions of Figures 4 and 5 confused me because I thought  
    that the stimuli on the right were “near” the training stimuli, when in fact the  
    arrangement of the stimuli matches that of Figure 3 and includes the training  
    stimuli on the left. I think the caption could make this clearer. I think one  
    reason for my confusion is the use of “near” and “far” labels in the test phases.  
    These terms are typically used to refer to physical distance from the training  
    stimuli. The far transfer test seems to be more of an explicit rule report test,  
    or a rule generation test (it seems to be interpreted in this way in the General  
    Discussion).  
  
    The wording of the test question in the Far Transfer Phase is slightly ambiguous.  
    Was this intentional? Perhaps more justification could be given for this question  
    (as opposed to something like “sort these marbles into 2 categories of your  
    choosing”).  
  
    The modelling results are informative in that the MLP proposed by Kurtz and Wetzel  
    (2021) fails to simulate the 2D results. Do the authors have any ideas on what  
    modifications could be made to the neural network to account for their results  
    (e.g., competition between dimensions/rules)?  
  
    In the GD, neural networks are pitted against explicit relational reasoning as  
    explanatory accounts. I didn’t see how performance in the far transfer task would  
    support the relational reasoning account over the neural network account, unless  
    learning in the neural network was assumed to be implicit. This argument needs  
    elaboration/refinement.  
  
    Minor  
    - The word “marbles” appears for the first time at the top of column 2 on page 2.  
    I don’t think this was explained. Related, what was the cover story given to  
    participants?  
    - The rationale for the control group was missing  
    - Would be good to see the proportion of participants using each strategy in a  
    table for easy comparison.  
  
----------------------------------------------------------------  
  
reviewer review (reviewer 2)  
score 1/5  
  
  Type of Submission  
  
    Cognitive Science, Psychology  
  
  The Review  
  
    (1) Submission's originality and significance  
  
    This study was an extension of Kurtz and Wetzel (2021) work with changing a bit  
    the diagnostic dimension used in the 2D condition and adding a far transfer  
    condition. The results are less reliable, as less than 30% of subjects in the 1D  
    and 2D conditions who  alternated their category predictions for the transfer  
    items in the way expected by the authors. Also, the computer simulation results  
    showed that the model MLP has difficulty simulating human's categorization  
    performance in the transfer phase. However, there is no clear explanation to it. A  
    future study which can provide mode solid evidence is needed.  
  
    (2) Technical soundness  
  
    The experimental design is appropriate. However, I am not sure why using computer  
    simulation instead of using a model to the fit the behavioral data. The neural  
    network model MLP is very simple and it would not surprise me, if it could not  
    make a good simulation. Also, this model could simulate the data in the 1D  
    condition. Why the model had different performance when simulating the data from  
    these two conditions can be addressed actually, since the authors still have 0.5  
    page to fill in.  
  
    (3) Theoretical merit  
  
    Relational category is an intriguing concept itself, that might reveal a different  
    way to represent categories from the conventional way (e.g., rule, exemplar, or  
    prototype). However, less than 1/3 of the participants in the 1D and 2D condition  
    actually showed alternation in the transfer phase. We might need more studies to  
    get a more reliable result.  
  
    (4) Breadth of interest for the broad Cognitive Science audience  
  
    As this study was an extension of Kurtz and Wetzel (2021), it might not attract  
    the attention of most Cognitive Science audience. Those who have read the paper of  
    Kurtz and Wetzel (2021) might be interested in the present study.  
  
  
    (5) Clarity of writing  
  
    The authors should be careful in editing. For example, the caption of Figure 2 was  
    shown before the figure of Figure 2. Also, on page 2, there is a strange section  
    Far Transfer Task, which looks like some editing mistake.  
  
----------------------------------------------------------------  
  
reviewer review (reviewer 3)  
score 3/5  
  
  Type of Submission  
  
    Cognitive Science  
  
  The Review  
  
    This study proposes that alternation can be learned as a basic category, and  
    presents some empirical and modeling results to support this idea. The empirical  
    results fairly convincingly demonstrate the phenomenon, showing that participants  
    most often learn an alternating category (rather than a simpler category based on  
    perceptual distance), and even generalize this category to a completely different  
    task with different stimuli. The main question that I'm left with is what we've  
    learned about categorization from these results. There are some limited modeling  
    results using a simple neural network model. But this model uses a sine function  
    as an activation in the hidden layer, and therefore builds in the notion of  
    alternation as an architectural primitive. This strikes me as being very ad-hoc  
    and unlikely to provide general insight into the processes that underlie human  
    category learning. This model also fails to capture the results in the 2d  
    condition. The authors also suggest that the phenomenon may be modeled as a kind  
    of relational category learning, but do not attempt to provide a concrete model to  
    instantiate this idea. In summary, I think the empirical results are interesting,  
    and convincingly demonstrate that alternation can be learned as a generalizable  
    category, but more work is needed to model and interpret these results relative to  
    more general theories of categorization.  
  
    Minor concerns:  
    - I found it very difficult to follow the description of the task setup. In the  
    'Stimuli' section, there is a reference to figure 2 that seems to imply it is a  
    depiction of the near transfer case, but then under figure 2 there appears to be a  
    caption describing the far transfer case. It's unclear if this is intended to go  
    with the figure or not. In general I think the clarity of this section could be  
    improved.  
    - The description of the background work (Kurtz & Wetzel (2021)) was also somewhat  
    difficult to follow without having read this previous paper. Perhaps it would be  
    useful to include a figure illustrating the previous work, which would then likely  
    make it easier to explain what is new in the present study.  
    - I really liked the method of visualizing the results for the near transfer  
    conditions (the heatmaps in figures 4 and 5). But in the previous work that this  
    paper builds on (Kurtz & Wetzel (2021)), there is also a visualization of the  
    individual subject results that I think would be informative to include here. It  
    would also be informative to see the individual subject results for the far  
    transfer task.  
    - In figure 6, the y axis should have a label.  
    - It seems that in the 2d task, the subjects picked an explanation based on one  
    dimension or the other (e.g. either the category split based on diameter, or the  
    alternation pattern based on angle), rather than learning a more complicated  
    composite function. I am curious what the author(s) think the explanation for this  
    might be, and how this behavior might be modeled?