Does active learning lead to better teaching of novel perceptual categories?

Oana Stanciu (stanciuo@ceu.edu)
Department of Cognitive Science, Central European University, Quellenstraße 51
Vienna, 1100 Austria

József Fiser (fiserj@ceu.edu)
Department of Cognitive Science, Central European University, Quellenstraße 51
Vienna, 1100 Austria

Abstract

To be efficient, both active learners and teachers need to be able to judge the relative usefulness of a piece of information for themselves or for their students, respectively. The current study assessed whether experience of active learning facilitates subsequent teaching from imperfect knowledge. Following a visual category learning task, dyads (N=40) of active and yoked passive learners taught (imagined) näive learners how to categorize the same visual stimuli by providing them with a small number of self-generated examples. Active learners narrowed down the possible categorization boundaries more than yoked learners. However, the active learning advantage was modest and limited to categories that were more difficult to learn and, overall, teachers were overly conservative, providing the least ambiguous category examples.

Keywords: teaching; active learning; yoked design

Introduction

Unquestionably, teaching by more knowledgeable and helpful others contributes to the efficiency of human learning and, more broadly, to our success as a species. Here, we focus on teaching by providing examples, which is particularly useful to teach concepts which are complex and difficult to verbalize, or for which explicit instruction does not aid learning (e.g., Rosedahl, Serota, & Ashby, 2021). While example giving is a mainstay of educational practice and the focus of the growing field of machine teaching, we know relatively little from an experimental standpoint about how humans explicitly generate examples, the extent to which they can intentionally modulate their sampling to meet different teaching goals, and how they can be scaffolded.

Generating teaching examples is a normatively hard problem, not least because teaching needs to be tailored to the learner's prior knowledge and their inferential process. Shafto, Goodman, and Griffiths (2014) proposed a solution by formalizing a rational-agent model of teaching as a recursive process: the teacher chooses examples for the learner such that the learner is most likely to adopt the hypothesis-to-be-taught, and the learner makes Bayes-rational inferences based on the examples received under the assumption that the teacher selected them as described above. The optimal (minimally sufficient) teaching set is converged upon by iteration.

Beyond the useful simplifying assumption of rational learners, another straightforward way to facilitate building a good learner model for teaching is for the teacher to first take the role of the learner (Stanciu, Lengyel, & Fiser, 2019), given that the ability to mentalize and take the perspective of the learner is prerequisite for good teaching (Bass, Shafto, & Gopnik, 2017). Being an active learner, that is, engaging in tasks where the learner can exert control over the learning curriculum (e.g., choosing which stimuli to get feedback on), should be the most beneficial to subsequent teaching since it would provide experience determining the potential usefulness of examples. The similarity between active learning and teaching is apparent when considering that the goal of active learners (who, unlike teachers, do not have access to the target hypothesis) is to sample their environment to maximize their expected information gain in light of their prior knowledge and the hypotheses that they wish to test. Based on the similarity at the computational level (see Yang and Shafto (2017) for a formalization of active learning as self-teaching), it may be the case that they rely on shared abilities which support the efficient sampling of data. As such, we expect that active learning experience will facilitate teaching (above and beyond passive forms of learning), even in the absence of explicit transfer of knowledge.

The current experiment tested this hypothesis by comparing the teaching performance of participants who learned how to categorize perceptual stimuli actively (i.e., by designing the stimuli they wanted to see labelled) versus yoked learners (i.e., learners who passively saw the labelled data selected by the the active learners). Expanding on previous literature (e.g., Avrahami et al., 1997), to reflect some of the complexity of teaching by giving examples in ecological contexts, learning was extended in time, the teacher did not have perfect knowledge of the hypothesis to be taught, and had to generate rather than select teaching examples. To titrate the difficulty of the task, two categorization rules were used, which also differed in their verbalizability (Rosedahl & Ashby, 2021).

The learning task was adapted from Markant and Gureckis (2014), who showed that active selection of training data improves categorization of two-dimensional visual stimuli with rule-based (RB) and information-integration (II) category structures. Both category boundaries were deterministic, but differed in the number of features one which they were based: one feature for RB and two for II. The active learning advantage was explained by Markant and Gureckis (2014) as a consequence of the fact that active learners generate individual hypotheses sequentially and choose the most informative queries to test their current hypothesis.

In L. K. Samuelson, S. L. Frank, M. Toneva, A. Mackey, & E. Hazeltine (Eds.), *Proceedings of the 46th Annual Conference of the Cognitive Science Society.*

We speculated that the advantage of active over yoked learners in teaching may be larger for the II category structure, as judging the informativeness is potentially more difficult when two features need to be taken into account. Further, previous work has showed that explicit written and verbal instruction did not improve performance in the II category learning task, but did so for RB category structures (Rosedahl et al., 2021), suggesting that teaching by example may be particularly well suited for this type of category structure.

Methods

Participants

80 adult participants (61 female, M_{age} = 26.67 years, range = 18 - 50 years old), predominantly university students, were recruited for a lab-based experiment in Budapest, Hungary [1]. Participants either volunteered their time in return for course credit, or received vouchers or monetary compensation (approximately €4). Ethical approval was obtained from EPKEB in Hungary.

Tasks

Category learning task The category learning task, including the cover story, was adapted from Markant and Gureckis (2014). Participants were told that they would see depictions of loop antennas that received one of two television channels. The antennas (see Figure 1) were visually represented by circles varying along two dimensions: the circle radius and the angle of a central diameter line. Which channel was received depended on how the antennas looked and the participants' task was to distinguish Channel 1 receiving antennas from Channel 2 receiving antennas. Unknown to the participants, the categorization boundary was either one-dimensional (based on either orientation or radius length, *RB structure*) or two-dimensional (*II structure*). Participants learned about the antennas either (*actively*) by designing their own and checking which channel they received, or (*passively*) by being shown antennas alongside the channel received.

Active learning trials started with the presentation of a randomly generated stimulus. The participant then modified the size of the circle and/or the orientation of the diameter line by moving the mouse horizontally while pressing one of two keyboard keys. Once the participant was satisfied with the designed stimulus, they could check the label by making a mouse click. The stimulus features could only be changed one at a time. Participants were required to manipulate at least one stimulus dimension to see the category label. There was no time limit for making alterations to the stimuli. The stimulus and category label were then presented together for 1,500 ms.

Crucially, the stimuli shown to **yoked** learners were the exact stimuli designed by their paired active learner, and were presented in the exact same order. Trials started with a brief fixation cross (250 ms), followed by the stimulus, which was shown for 250 ms on its own before the channel label was added. The stimulus and label were jointly presented until the participant pressed a button corresponding to the label shown on the screento ensure that participants attended the stimuli.

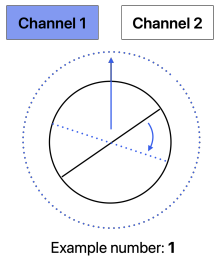


Figure 1: Illustration of a teaching task trial. The participant first chooses the category by clicking on one of the channel buttons and then produces the example by modifying a random stimulus.

Test trials were identical across the two learning conditions. A stimulus was presented on the screen and the participant had to press one of two buttons corresponding to its category. Participants only received aggregated feedback at the end of every block. Test stimuli were sampled uniformly from the quadrants of the stimulus space to avoid biasing participants' category representations and to ensure chance performance was 50%.

Teaching task Participants designed example stimuli using the same procedure that active learners used to generate queries. To ensure that yoked learners were equally competent at producing samples, after familiarization with the setup, they performed (at least four trials) of an additional practice task. Specifically, participants had to manipulate stimuli until they perfectly matched a target antenna presented on the screen.

Participants were instructed to teach another (fictitious) participant, who they were told was yet to take part in our experiment, which antennas receive Channel 1 and Channel 2. It was stressed that these participants were naïve and that they will complete the same categorization test. Participants were then told that the only constraint for the teaching task was that they can only provide a maximum of six antenna examples. They were encouraged to give examples that would be as helpful as possible to the learner.

Participants first had to choose the channel their example antenna received by clicking on one of two buttons (see Figure 1). Then, a randomly generated antenna was drawn on the screen and they could manipulate it to design their intended example. An example counter was always presented in the bottom corner of the screen to ensure participants knew how many examples they had selected so far. Example selection was unspeeded.

At the end of the task, participants were prompted to type answers to open-ended questions about their teaching strategies, and how they would have taught another participant if

[1] An additional participant was excluded due to lack of task compliance.

they could verbally communicate to them.

Stimuli and materials

The stimuli were circles defined by the size of the radius and the orientation of the diameter, two features shown to be perceived as independently by Nosofsky (1989). The same range of circle radius and orientation (90°, to avoid use of circular variables) was used for every participant, but the minimum radius and angle values were pseudo-randomly sampled.The category boundary always halved the stimulus space. Since every participant had a different boundary in perceptual space, stimuli are rescaled to an abstract stimulus space for analyses and illustrations.

All variables governing the presentation of the stimuli were counterbalanced: the feature relevant for RB classification, the diagonal used for the II categorization, the mapping between keyboard buttons and features, and between the mouse movement direction and the direction of changes in the stimuli.

The number of teaching examples was fixed to six to ensure a meaningful comparison between participants. A relatively low number of examples (but sufficient to optimally describe the boundaries) was used to stress the importance of selecting good examples, but also because pedagogical and random sampling become harder to distinguish with larger numbers of samples.

The task was implemented in *PsychoPy*3 (Peirce et al., 2019).

Design and Procedure

Participants completed a category learning task followed by a teaching task. Importantly, participants were not informed that they will be asked to teach before starting the learning task. The learning task had a 2 (learning condition: active, yoked passive) by 2 (category structure: RB, II) between participants design. Participants were pseudo-randomly assigned to conditions as each passive participant was yoked to an active learner (creating 40 dyads comprising an active and a yoked passive learner). There were eight blocks, each consisting of 16 learning trials immediately followed by 32 test trials. On average, the task took 44 minutes and 21 minutes for active learning and yoked participants, respectively.

Data analysis

Category learning task Active and yoked learners' test accuracy was compared using both between-group and within-dyad comparisons. The sampling behavior of the active learners was quantified by the distance from their queries to the true boundary, as well as to their subjective boundaries, fitted for every block using a decision-bound model. Participants were assumed to provide probabilistic category membership responses for a stimulus as a function of its location relative to a linear boundary traversing the two-dimensional stimulus space. The likelihood of one stimulus being categorized as 'Channel 1' is given by Equation 1 where x is the stimulus as defined by the two perceptual dimensions, θ is the angle of the linear boundary, b is the orthogonal distance from the center of the space to the boundary, and σ expresses how deterministic the category responses are (high values indicating more deterministic responses).

$$P(x^{trial} = \text{CH1}|\theta, b, \sigma) = \frac{1}{1+\exp(-\sigma(x_1^{trial} \cdot \cos(\theta) + x_2^{trial} \cdot \sin(\theta) - b))} \quad (1)$$

Reference to subjective boundaries is especially relevant in the II condition, where many participants did not converge to the true boundary. For every participant, the decision-bound model was compared (using the Aikaike Information Criterion, AIC) to a random-response model in which participants categorized stimuli based on a fixed probability, uninformed by the distance of stimuli to the boundary. Participants who were not better fitted by the decision-bound model ($\approx 10\%$) were eliminated from the analysis.

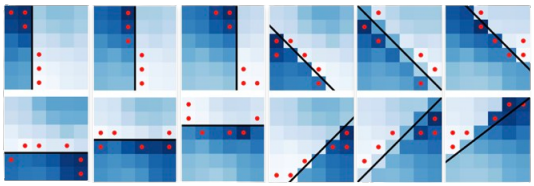
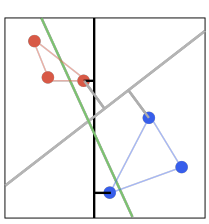


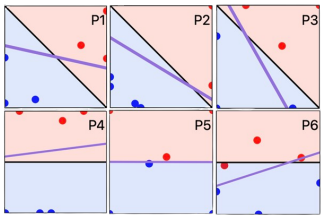
Figure 2: Teaching sets selected by the pedagogical model (red) for each hypothesis (black). Heatmaps illustrate the likelihood of choosing examples at different locations in the stimulus space.

Teaching performance quantification and predictions As a preliminary check, we tested whether the chosen examples in the different conditions significantly departed from random uniform sampling across the stimulus space using Kolmogorov-Smirnoff tests (KS). More meaningfully, the individual decision-bound model fits were used to compute the likelihood of examples under random sampling assuming the decision bound model (strong sampling). This intuitively predicts that an example is more likely to be selected if its orthogonal distance to the category boundary is larger, but leads to a large number of equiprobable teaching tests. Using random sampling to teach would provide accurate examples for learners, but would be inefficient. To provide a normative benchmark, predictions were also generated from the Shafto et al. (2014) iterative pedagogical sampling model (see Figure 2) using a simplified, tractable set-up. This model predicts that teaching examples should be located closely around the boundary if teachers are certain about the location of the boundary.

Descriptive metrics of teaching performance were computed to provide an intuitive understanding of the different (likely implicit) teaching "strategies" used by participants (see Figure 3) and, in an exploratory analysis, tested within-dyad differences. First, whether the learner is näive or not (i.e., assumes they are taught or not), example sets which are compatible with fewer boundaries are more useful. Using a fine grid over the bivariate space of parameters governing the



(a) Learner inference



(b) Example participants

Figure 3: (a) Possible boundary inferences of a learner based on a set of labeled examples: The black and gray (but not the green) boundaries perfectly separate the examples. The black line minimizes the orthogonal distance between the 2 closest examples of a different category. The gray line maximizes it. A näive learner will exclude the inconsistent boundary, but would not have any reason to prefer one consistent boundary over another. Learners who know they are taught may believe that the black boundary is more likely because it minimizes the distance between categories. (b) Teaching samples (dots) chosen by 6 participants alongside their fitted subjective boundaries (purple line). Dot colors represent the category labels chosen by the participant, area shading indicates ground truth. The orthogonal distances from the examples to the subjective boundaries of participants 6 and 7 are very small, but a näive participant observing P6 examples might infer a diagonal boundary leading them to wrongly classify a large area of the stimulus space. The potential for misclassification based on P7 is low.

linear boundaries, we calculated the proportion of discretized boundaries consistent with the teaching set (i.e., separated examples from the two conditions, Figure 3). A related strategy (but which does not assume linear classification) is to give examples which cover the entire (or as much as possible) of the category space. To measure this, we calculated and summed the area of the polygons inscribed by the examples labeled as the same category (generally, triangles), removing intersecting areas.

Another proxy for fewer boundaries being compatible with the teaching set is to provide at least two examples that are as close as possible to the subjective boundary. To measure this, the 'boundary distance' was computed as the minimal orthogonal distance of the closest examples on either side of the teacher's subjective boundary. It is also possible, that learners viewing pedagogically produced sets, may reason that, if multiple boundaries are compatible, the one minimizing the distance between example of different categories may be the intended one (e.g., the black line in Figure 3).

Further, if teachers are sensitive to their own limitations as learners, there should be a negative correlation between how deterministic they were in the categorization (σ fitted in the last test block of the experiment) and the boundary distance measure described above. In other words, poor learners, who were still uncertain about the location of the boundary by the end of the training, were expected to choose teaching examples farther apart to avoid mislabeling them, and therefore, misguiding the learner.

Lastly, especially when teaching RB categories, teachers can highlight the feature relevant for classification by manipulating variability, namely, by keeping one feature constant while introducing maximal variance in the irrelevant (or orthogonal) feature. The ratio of the variances of the two features of examples was used to measure this.

The distribution of each metrics in the sample was compared to random expectations (i.e., when choosing examples uniformly at random from the true category), and within-dyad differences were tested. Since active learners were expected to overperformed yoked learners in categorization test by the end of the task, this raised the concern that any within-dyad differences in teaching performance stem solely from the yoked learners' more uncertain/poorer category representation. Within-dyad differences in teaching were regressed on within-dyad differences in accuracy in the final test block.

Results

Categorization Performance

Results were highly consistent with the findings of Markant and Gureckis (2014). Participants learning the RB structure outperformed participants learning the II structure at test(see Figure 4). Active learners, regardless of condition, were more likely to be correct in the categorization test than yoked learners. A 2x2 between participants ANOVA resulted in significant main effects of learning mode, $F(1,76) = 7.04, p < .01$, and category structure, $F(1,76) = 18.64, p < .001$, without a significant interaction, $F(1,76) = .01, p = .99$. These differences remained significant in the last test block of the experiment. Further, within-dyad differences between active and yoked learners were statistically significant in paired t-tests, $t(19) = 3.04, p = .01, BF_{alt} = 7.13$, for the RB structure, and, $t(19) = 2.82, p = .01, BF_{alt} = 4.75$, for the II structure.

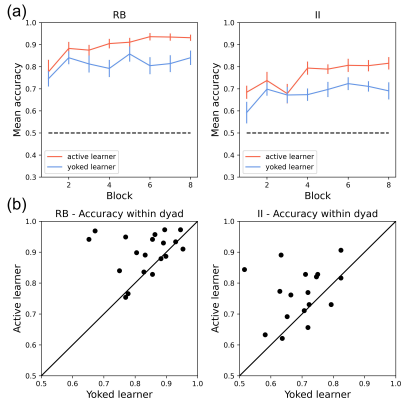


Figure 4: (a) Average categorization accuracy (+/-SE) by experimental condition and block (Upper panel). Using a unidimensional rule in the II condition results in about 75% average accuracy. (b) Within dyad accuracy differences. Each dot is a dyad. Axes display the proportion of correct responses.

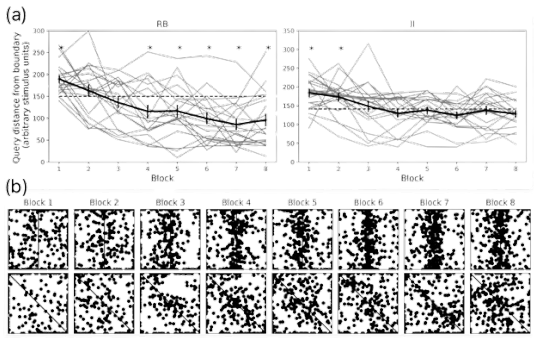


Figure 5: (a) Average query distance to boundary (+/- SE) for the group (black) and individual participants (gray). The dashed line is the expected query distance under random sampling for each corresponding category structure and asterisks mark blocks that are significantly different in one-sample two-tailed t-tests ($\alpha < .05$). (b) Queries made by all participants across the 8 active learning blocks of the RB (Upper panel) and II (Lower panel) categories. Each dot corresponds to the angle and radius defining an antenna. Stimuli have been rotated to align true boundaries across participants.

For the RB category structure, subjective boundaries converged by the end of the task to the true boundary, for both active and yoked participants. In the II structure, subjective boundaries tended to be axis-aligned, especially in the early blocks. For yoked participants learning the II structure, there was no discernible convergence pattern by the end of the task (65% were better fit by a RB boundary, based on AIC), but a larger proportion of active learners in the II condition were fitted by boundaries that were not axis-aligned (85%) and relatively close to the true boundary.

Active learning performance

All active learners started by making average queries that were farther away from the boundary than can be expected based on a random sampling strategy, consistent with an early exploration of the extremes of the stimulus space (see Figure 5). While participants in the RB condition made average queries that were lower than the random-sampling expectation starting roughly from the middle of the task, this never happened for participants in the II condition (see Figure 5). In the final block of the task, RB participants made queries well below chance level, $t(19) = -3.90, p < .001, BF_{alt} = 37.80$ (2-tailed), but not II participants, $t(19) = -1.56, p = .13, BF_{null} = 1.52$.

Active learners who chose samples closer to the boundary performed better at test, both in the RB ($r(18) = -.57, p < .01$) and II ($r(18) = -.64, p < .01$) conditions. On the other hand, there was no significant correlation between the test accuracy of passive learners and the distance from boundary of the stimuli they saw ($RB: r(18) = -.08, p = .72; II: r(18) = -.20, p = .39$). The correlation observed for active learners was statistically different from that observed for yoked learners according to a Fisher z-transformation ($z = -1.93, p = .05$).

Teaching Performance

What examples did teachers choose? Teachers overwhelmingly labelled examples correctly (90% of time) and generally chose to give an equal number of examples from the two categories (77.50%).

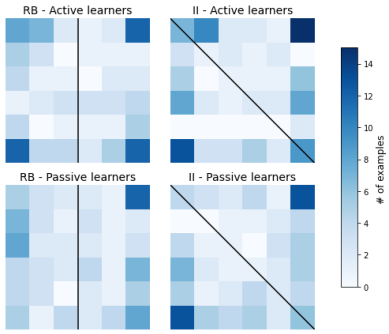


Figure 6: Distribution of teaching examples across the stimulus space (pooled across all participants). Each cell contains about 3% of samples under uniform sampling of the stimulus space.

The extremes of the stimulus space (especially corners) were oversampled, but locations close to the boundary were not (see Figure 6), and as such participants were best fit by strong sampling from the decision-bound model. The overrepresentation of extreme samples was expected for II categories given the relatively high uncertainty about the boundary, but was surprising for the RB category since most participants were very precise in the last categorization test and were highly deterministic in their labelling. Visual inspection also did not reveal consistent order patterns across samples such as curriculum teaching (i.e., starting with easy examples far from the boundary and increasing difficulty gradually by providing examples closer to the boundary). The RB irrelevant feature samples produced by yoked, $D = .22, p = .34$, but not by active learners, $D = .36, p = .02$, were consistent with uniform sampling.

Were active learners better teachers than yoked learners? As seen in Figure 7, active learners chose example sets which left fewer compatible boundaries than passive learners in the II condition, $t(15) = -2.26, p = .04, BF_{alt} = 1.82$, but not in the RB condition, $t(15) = -.72, p = .48, BF_{null} = 3.12$. There was no difference in the number of consistent boundaries between the two category structures, $t(62) = -.36, p = .72, BF_{null} = 3.70$.

Active learners selected examples that inscribed a significantly larger area of the stimulus space than yoked learners in the II condition, $t(17) = 3.66, p < .001, BF_{alt} = 21.13$, but not in the RB condition, $t(17) = 1.87, p = .08, BF_{alt} = 1.02$.

There were no significant within-dyad differences in min-

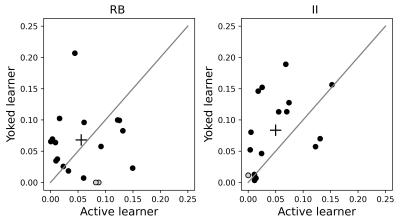


Figure 7: Proportion of boundaries compatible with the teaching examples. Each dot is a dyad (in grey dyads were excluded from analysis because one participant did not choose a linearly separable example set). Crosses mark sample means.

imum example to boundary distances in either condition[2], RB: $t(17) = -.72, p = .48, BF_{null} = 3.11$, II: $t(15) = .02, p = .98, BF_{null} = 3.91$. There was also no evidence that participants introduced more variance in relevant vs irrelevant(RB)/orthogonal(II) feature across any of the groups, all ps$>$.5.

The fitted σ correlated negatively with the average distance from the examples to the boundary for teachers who were active learners, but it was not significant, $RB: r(16) = -.44, p = .07, BF_{alt} = 1.79; II: r(16) = -.12, p = .62, BF_{null} = 1.85$. The correlation in the yoked groups was $RB: r(16) = .02, p = .95, BF_{alt} = 2.01; II: r(16) = .05, p = .85, BF_{alt} = 1.99$.

Did teaching differences within-dyads rely on within-dyad categorization differences? The difference in the number of consistent boundaries for dyad members correlated negatively (as predicted), but not significantly, with accuracy differences for the II category, $r(14) = -.46, p = .07$. There was no correlation in the RB category, $r(14) = .16, p = .55$.

Are inter-individual differences in active learning predictive of teaching performance? Participants who made queries further away from their subjective boundary tended to be worse teachers in the II condition (as indexed by the proportion of compatible boundaries), $II: r(15) = .47; p = .05, BF_{alt} = 2.03$, but not in the RB condition ($RB: r(15) = -.16; p > .05$). Accuracy correlated positively, but non-significantly, with teaching performance ($RB: r(15) = -.39, p = .10$; $II: r(15) = -.44, p = .07$). The partial correlation between teaching and active learning performance, controlling for accuracy at the end of the task, was $r(15) = .32, p = .22$, also non-significant.

Discussion

The active learning advantage found by Markant and Gureckis (2014) in the categorization task was successfully replicated. Did the active learning experience also improve selection of examples for teaching? Active learners, compared to yoked learners, generated teaching sets which where compatible with a smaller proportion of linear boundaries and inscribed a larger area of the stimulus space, but only for II category structures. There were no differences in how close to the boundary teachers placed their examples or the variability of the examples across features, leading to a mixed pattern of results. Given also that the metrics described here were exploratory and the sample sizes were relatively small, replication of the findings would be needed before drawing further conclusions.

On the other hand, what is clear is that across the board, participants tended to oversample the edges of the stimulus space, providing unambiguous examples. However, teachers were very conservative in terms of the distance of the samples to the boundary, beyond what would be warranted by the noise in their categorization decisions, at least for RB categories, where they had very good categorization performance and were given explicit feedback regarding this. We speculated that the active learning advantage might be larger for II compared to RB categories because teaching II categories is more difficult, the poor performance on the RB task, raises new questions about the origin of the category differences.

Further, we did not find the predicted correlation between how deterministic participants were in labeling of stimuli during the final test block (indicative of uncertainty around the boundary location) and how close they placed examples to their subjective boundary. However, better active learners of II categories (who were designing samples closer to the boundary), were found to generate teaching sets which eliminated more possible boundaries. It is also interesting to note that despite following a distinctive pattern in their active learning of starting with the extremes of the stimulus space and finishing by choosing examples very close to the boundary, participants did not reproduce patterns consistent with such curriculum learning in their teaching, as observed in Khan, Zhu, and Mutlu (2017).

All in all, it seems that the participants in our sample were not efficient teachers. Of course, the best test of their teaching skills would have been to present human learners with the samples they produced. However, we think it is unlikely that the small differences observed between teaching sets would translate into meaningful differences in categorization accuracy or category acquisition time for new learners. Follow-up studies should address whether poor teaching performance on this task (and in contrast with other work e.g., Shafto et al., 2014) is related to the participant's uncertainty around the true boundary (e.g., by automatically labelling examples as the participant designs them during teaching and eliminating uncertainty around the categorization) or the fact that teachers had to construct examples using fine grained continuous features as opposed to selected sets of individuated category members (e.g., presented on individual cards), similarly to the generation/selection gap in question-asking (e.g., Rothe, Lake, & Gureckis, 2018).

Lastly, it would be interesting to see if more meaningful teaching examples would be elicited in a truly interactive settings - where teachers can observe and adapt to the impact of their examples on learners.

[2]Data from 2 participants who only selected examples from one category was excluded.

References

Avrahami, J., Kareev, Y., Bogot, Y., Caspi, R., Dunaevsky, S., & Lerner, S. (1997). Teaching by examples: Implications for the process of category acquisition. *The Quarterly Journal of Experimental Psychology Section A*, *50*(3), 586–606. Retrieved from https://doi.org/10.1080/713755719 (_eprint: https://doi.org/10.1080/713755719) doi: 10.1080/713755719

Bass, I., Shafto, P., & Gopnik, A. (2017). I know what you need to know: Children's developing theory of mind and pedagogical evidence selection. In *Proceedings of the 39th annual conference of the cognitive science society* (p. 6).

Khan, F., Zhu, X., & Mutlu, B. (2017). How do humans teach: On curriculum learning and teaching dimension. *Advances in Neural Information Processing Systems 30 (NIPS 2017)*.

Markant, D. B., & Gureckis, T. M. (2014). Is it better to select or to receive? Learning via active and passive hypothesis testing. *Journal of Experimental Psychology: General*, *143*(1), 94–122.

Nosofsky, R. M. (1989). Further tests of an exemplar-similarity approach to relating identification and categorization. *Perception & psychophysics*, *45*(4), 279-290. Retrieved from https://doi.org/10.3758/bf03204942

Peirce, J., Gray, J. R., Simpson, S., MacAskill, M., Höchenberger, R., Sogo, H., ... Lindeløv, J. K. (2019). PsychoPy2: Experiments in behavior made easy. , *51*(1), 195–203. (Place: Germany Publisher: Springer) doi: 10.3758/s13428-018-01193-y

Rosedahl, L. A., & Ashby, F. G. (2021). Linear separability, irrelevant variability, and categorization difficulty. *Journal of experimental psychology. Learning, memory, and cognition*, *18*.

Rosedahl, L. A., Serota, R., & Ashby, F. G. (2021). When instructions don't help: Knowing the optimal strategy facilitates rule-based but not information-integration category learning. *Journal of Experimental Psychology: Human Perception and Performance*, *47*(9), 1226–1236. (Place: US Publisher: American Psychological Association) doi: 10.1037/xhp0000940

Rothe, A., Lake, B. M., & Gureckis, T. M. (2018, March). Do People Ask Good Questions? *Computational Brain & Behavior*, *1*(1), 69–89. Retrieved from https://doi.org/10.1007/s42113-018-0005-5 doi: 10.1007/s42113-018-0005-5

Shafto, P., Goodman, N. D., & Griffiths, T. L. (2014). A rational account of pedagogical reasoning: Teaching by, and learning from, examples. *Cognitive Psychology*, *71*, 55–89.

Stanciu, O., Lengyel, M., & Fiser, J. (2019). To teach better, learn first. *Proceedings of the 41st Annual Meeting of the Cognitive Science Society*.

Yang, S. C.-H., & Shafto, P. (2017). Teaching Versus Active Learning: A Computational Analysis of Conditions that Affect Learning. In *Proceedings of the 39th annual conference of the cognitive science society.*