

**THE FEASIBILITY OF
HUMAN TEACHING BY SAMPLING
AND THE CHALLENGES OF LEARNING FROM TEACHERS**

By

Oana Stanciu

Submitted to

Central European University

Department of Cognitive Science

*In partial fulfillment of the requirements for the degree of Doctor of
Philosophy in Cognitive Science*

Primary Supervisor: József Fiser

Secondary Supervisor: Máté Lengyel

Budapest, Hungary

2022

Declaration of Authorship

I hereby declare that this submission is my own work and to the best of my knowledge it contains no materials previously published or written by another person, or which have been accepted for the award of any other degree or diploma at Central European University or any other educational institution, except where due acknowledgment is made in the form of bibliographical reference.

Oana Stanciu

Abstract

From a young age, humans efficiently utilize sparse observable samples to make reliable inferences about a highly complex, stochastic and ever-changing world. Undoubtedly, the social grounding of human inductive learning, especially through teaching by more knowledgeable and helpful others, contributes greatly to the success of human learners. However, while there is tremendous interest in the development of intelligent tutoring systems for educational applications, and significant theoretical and applied advancements have been made in the burgeoning field of machine teaching, experimental work in cognitive science has focused almost exclusively on investigating the behaviors of learners, largely overlooking teachers.

This thesis investigated one of the main ways in which humans teach, both in formal and informal settings: by offering examples. Since explicitly generating samples for the purpose of teaching others is undeniably a normatively hard problem, we explored possible limitations faced by teachers due to the abstraction and complexity of the task (Chapter 2) and having a good model of the learner (Chapter 3). From the learner's perspective, we examined whether learners could effectively learn from sampled data (Chapter 2) and assess imperfect teachers (Chapter 4).

In Chapter 2, replications and extensions of two teaching games (prototype and category boundary teaching) lead to diverging results in terms of the optimality of teachers, which we attributed to differences in the level of abstraction of the task and the complexity of the stimuli. In turn, learners were also limited in their ability to effectively adapt to the data generative process, specifically when producing estimates based on autocorrelated samples, a known feature of samples produced by humans.

In Chapter 3, we proposed that teachers can overcome one of the more challenging aspects of teaching - building an adequate model of how learners make inferences - by engaging in the experience of learning, and especially active learning (given the computational similarity). We present evidence that prior active learning facilitates teaching from two different category teaching experiments.

Lastly, in Chapter 4, we proposed that confidence may be a useful pedagogical signal. For

instance, explicit confidence statements could allow teachers to monitor the progress of their students, and help learners to choose better teachers. Focusing on the learners' perspective, a first experiment found, in agreement with previous work, that humans preferred more informative and better calibrated advisers as future collaborators (even when compared with overconfident advisers). Interestingly, a second experiment showed a dissociation between partner preferences and decision making. Specifically, learners did not optimally use information about the relative metacognitive skills of informants to improve their decision making.

In light of the results presented, we suggest that while rational pedagogical models can be a useful computational-level description of teaching solutions in some limited domains, they are unlikely to provide a close account of flexible, on the fly teaching behaviors without significant modifications that account for the important limitations facing teachers.

Acknowledgements

First, I would like to thank my supervisor, József Fiser, for the the teaching moments, the many conversations over science or wine (or both), and the continued support and encouragement. I also owe many thanks to my second advisor, Máté Lengyel, for the always astute feedback and invaluable know-how about how to more closely approximate principled experimentation.

I am grateful to all Vision Lab members for the feedback and the fun memories (and the free dog sitting), especially to those who have been there from the start of my PhD: Sara, Bozsi, Gabor, Marci, Adam, Benó.

To the many friends I made in the department across the years, to the collaborators and collaborating friends - I am happy to have been on this journey together and look forward to the next one!

Mia, I am not overstating when I say I can't (or wish to) imagine these past years in Budapest without you! Jackie, thanks for always being my family away from home!

I thank my family for their support and for putting up with me never having a convincing answer for when I'm expected to graduate or what exactly it is that I am working on.

Last but not least, I would like to thank the hundreds of people who volunteered their time to participate in the experiments I conducted during these years.

Table of Contents

| | | |
|----------|--|-----------|
| 1 | Introduction | 1 |
| 1.1 | Brief overview: How teachers help learners | 1 |
| 1.2 | Motivation | 3 |
| 1.3 | Are teachers optimal? Setting the problem | 5 |
| 1.4 | Optimal teachers in the real world | 9 |
| 1.4.1 | Learning about others | 9 |
| 1.4.2 | Representation of the task environment and hypothesis generation . . . | 14 |
| 1.4.3 | Teaching is extended in time | 15 |
| 1.4.4 | Learners are active | 16 |
| 1.4.5 | Redundancies can be optimal for bounded learners | 16 |
| 1.5 | Conclusions and further directions | 19 |
| 2 | Sampling for teaching and learning | 21 |
| 2.1 | Introduction | 21 |
| 2.2 | Study 1: Teachers | 22 |
| 2.2.1 | Replication of Shafto et al. (2014): Prototype teaching | 24 |
| 2.2.2 | Replication of Khan et al. (2011): Boundary teaching | 28 |
| 2.2.3 | Conclusion | 32 |
| 2.3 | Study 2: Learners | 35 |
| 2.3.1 | Overview of experiments | 35 |
| 2.3.2 | Defining optimal performance | 37 |

| | | |
|----------|---|-----------|
| 2.3.3 | Validation of data generation | 39 |
| 2.3.4 | Defining near -optimal performance | 40 |
| 2.3.5 | Performance improvement conferred by the optimal strategy | 44 |
| 2.3.6 | Experiment 1 | 46 |
| 2.3.7 | Follow-up: Univariate stimuli | 52 |
| 2.3.8 | Follow-up: Prediction task | 52 |
| 2.3.9 | Experiment 2 | 54 |
| 2.3.10 | Follow-up: Memory task | 55 |
| 2.3.11 | Conclusion | 57 |
| 2.4 | General Discussion | 60 |
| 2.5 | Supplementary Information | 63 |
| 2.5.1 | Study 2: Additional figures | 63 |
| 3 | Actively Learning How to Teach | 67 |
| 3.1 | Introduction | 67 |
| 3.2 | Study motivation | 72 |
| 3.3 | Experiment 1 | 75 |
| 3.3.1 | Introduction | 75 |
| 3.3.2 | Methods | 75 |
| 3.3.3 | Results | 81 |
| 3.3.4 | Conclusion | 86 |
| 3.4 | Experiment 2 | 88 |
| 3.4.1 | Introduction | 88 |
| 3.4.2 | Methods | 91 |
| 3.4.3 | Results | 104 |
| 3.4.4 | Conclusion | 119 |
| 3.5 | General Discussion | 124 |
| 3.6 | Supplementary Information | 127 |

| | | |
|----------|--|------------|
| 3.6.1 | Pilot: Priors about boundary locations | 127 |
| 3.6.2 | Experiment 1: Additional analyses and figures | 130 |
| 3.6.3 | Experiment 2: Additional analyses and figures | 135 |
| 4 | Tracking others' confidence to select who to learn from | 149 |
| 4.1 | Introduction | 149 |
| 4.2 | Study motivation | 164 |
| 4.3 | Experiment 1 | 168 |
| 4.3.1 | Introduction | 168 |
| 4.3.2 | Methods | 171 |
| 4.3.3 | Predictions | 182 |
| 4.3.4 | Results | 183 |
| 4.3.5 | Discussion | 189 |
| 4.4 | Experiment 2 | 192 |
| 4.4.1 | Introduction | 192 |
| 4.4.2 | Methods | 194 |
| 4.4.3 | Results | 202 |
| 4.4.4 | Discussion | 206 |
| 4.5 | General Discussion | 208 |
| 4.6 | Supplementary Information | 210 |
| 4.6.1 | Experiment 1: Supplementary figures | 210 |
| 4.6.2 | Experiment 2: Supplementary figures | 213 |
| 4.6.3 | Experiment 2: Pilot | 219 |
| 5 | General Discussion | 221 |
| 5.1 | Conclusions | 221 |
| 5.2 | General limitations and further directions | 226 |
| | References | 229 |

List of Figures

| | | |
|------|--|----|
| 1.1 | Schematic representation of pedagogical reasoning from Shafto et al. (2014). | 2 |
| 2.1 | Empirical results of Shafto et al. (2014) (Fig. 7) along with model predictions. | 26 |
| 2.2 | Triplets of examples chosen by the teachers. L: large, M: medium, S: small. | 27 |
| 2.3 | Example stimulus array from the Khan et al. (2011) teaching task. Participants ranked images as a function of ‘graspability’ and then chose a boundary between graspable and non-graspable items. The image was reproduced from Khan et al. (2011). | 28 |
| 2.4 | The trajectory of the examples chosen by participants conforming to the curriculum learning strategy. The image was reproduced from Khan et al. (2011). | 28 |
| 2.5 | The trajectory of the examples chosen by participants. Circles represented example choices in the Unconstrained condition (the order of choice is down-up). Diamonds represent the examples selected in the Constrained condition. | 28 |
| 2.6 | Example image array from the teaching task. In this trial, images were sorted according to loudness. The gray dot marks the boundary chosen by the participant. In order to select examples for teaching, the participant had to drag either the red (not loud) or blue (loud) circles across an image, which will automatically mark it as an example by drawing a square of the corresponding color around it. | 29 |
| 2.7 | Teaching performance in replication of Khan et al. (2011). | 34 |
| 2.8 | Optimal weighting for a series of length ten as a function of the γ parameter which quantifies dependence. | 39 |
| 2.9 | The histograms of posterior samples over the correlation coefficient (upper) or standard deviation of the innovation (lower). Each subplot corresponds to one stimulated dataset of 300 trials, where each trial is a series of length 8. The red vertical lines correspond to the true value of the parameter. | 42 |
| 2.10 | Optimal weights derived using the observed covariance structure (‘Empirically optimal’ weights) of 1000 simulated datasets with sequences of length 8 and using 300 trials (upper) and 1200 trials (lower). | 43 |
| 2.11 | Comparison of uniform and optimal weighting strategies. | 45 |
| 2.12 | Sample end frame from the experiment demonstration. | 48 |
| 2.13 | Average weights. | 50 |
| 2.14 | Individual weights for the horizontal (manipulated) location alongside the sample average. | 51 |
| 2.15 | Examples of participant weights. | 52 |
| 2.16 | Average weights for estimation of the sequence center. | 53 |

| | | |
|------|--|----|
| 2.17 | Average prediction weights. | 53 |
| 2.18 | Sample stimuli for the estimation task. | 54 |
| 2.19 | Estimated weights. | 55 |
| 2.20 | Estimated weights. | 55 |
| 2.21 | Mean recall performance by condition. Vertical segments mark SEs. | 57 |
| 2.22 | Average estimation weights. | 58 |
| 2.23 | Relationship between serial effects in estimation weights and memory performance. | 59 |
| 2.24 | Estimation performance of Participant 1. The observed signed distances from arithmetic means of trials and the participant's estimates (red) against the distances between estimates and shuffled trial means. | 63 |
| 2.25 | True center for trials against means computed using the optimal strategy and uniform weights. | 64 |
| 2.26 | Variance explained (R^2) | 65 |
| 2.27 | Predictions based on optimal and uniform weighting (Left); Ground truth relative to optimal and uniform weighting (Right). | 65 |
| 2.28 | AIC values for the models fit in Experiment 1. Each row corresponds to a participant. | 66 |
| 3.1 | Example image array from the teaching task. In this trial, food items were sorted from left to right in ascending order of their vitamin B content. The black vertical bar represents the daily recommended dose of vitamin B, which is the boundary the participant had to teach. In this case, the participant clicked on the two images closest to the boundary, which were automatically labelled. | 77 |
| 3.2 | Teaching and learning performance across the three conditions. Each dot represents the information gain for one participant, averaged across the three trials of each task. Crosses represent the 95% confidence intervals for the group means. Dotted lines represent the expected mean information gain from teaching as a function of expected information gain. Gray lines mark chance performance (see Supplementary Information for details). The maximum information gain for the task is 2.81 bits. The asterisks mark significance levels in independent t-tests (* $p < .05$, ** $p < .01$). | 82 |
| 3.3 | Teaching performance for the active-passive learning dyads. Each dot represents the information gain from teaching for one dyad. In dyads situated under the diagonal identity line, the active learner was the better teacher. A small Gaussian scatter was applied to make overlapping dots visible. | 84 |
| 3.4 | The difference in teaching information gain within dyads of active and passive learners as a function of the expected information gain for (active) learning. The fitted OLS regression line is shown alongside its 95% confidence bound. Learning performance was not a significant regressor of the difference in teaching, $\beta = -.86$, $p = .07$. The predicted within-dyad teaching difference was .60 bits, $p = .03$, at a one bit expected learning entropy and decreased to zero for dyads with high expected information gain. | 85 |
| 3.5 | Average group task performance (+/- SEM) broken down by trial number. | 86 |

| | | |
|------|---|-----|
| 3.6 | Random sampling of stimuli to offer as examples for the four boundaries used in the task. The farther from the boundary an example is, the more likely to be chosen as a category example. | 97 |
| 3.7 | Predictions from the pedagogical model for the likelihood of choosing examples at different locations in the stimulus space. The most probable example sets are overlaid in red and the hypotheses are drawn in black. | 99 |
| 3.8 | Predictions of the pedagogical model for teaching only RB categories. Red dots represent the most likely example set under pedagogical sampling. | 100 |
| 3.9 | Teaching samples (dots in the two dimensional stimulus space) offered by two experiment participants in the RB condition. Purple lines represent their subjective boundary inferred based on the last test block. Dot colors represent the category labels chosen by the participant and area colors represent the ground truth. | 101 |
| 3.10 | Overall accuracy in categorization. Bars represent standard error of the mean. | 105 |
| 3.11 | Average categorization accuracy by experimental condition and block order (Upper panel). Vertical lines represent the standard error of the mean. Using a unidimensional rule in the II condition would result in about 75% average accuracy. Lower panel shows within dyad categorization performance differences. Each dot represents a dyad. All axes display the proportion of correct responses. | 106 |
| 3.12 | Fitted boundaries for all participants (purple) against true boundaries (black). Stimulus spaces were rotated in order to be able to overlay boundaries of all participants. | 107 |
| 3.13 | Average query distance to boundary for the group (black) and individual participants (gray). Bars represent standard errors of the mean. The dashed line is the expected query distance under random sampling for each corresponding category structure. Asterisks are displayed above blocks where the average query distance differed from the chance significantly in a one-sample two-tailed t-test ($\alpha < .05$). | 108 |
| 3.14 | Queries made by all participants across the 8 active learning blocks. Each dot corresponds to the angle and radius size that defined a stimulus. Stimuli have been rotated such that the true boundaries are aligned for all participants. | 109 |
| 3.15 | Scatterplot of the average query distance from the true boundary against overall categorization performance (of active learners or paired yoked learners). Each dot is a participant. Correlations are significant for active learners (upper plots), but not between an active learner's categorization accuracy and their paired active learner's query distance (lower plots). | 109 |
| 3.16 | Frequency of teaching examples across the stimulus space (pooled across all participants). Color intensity corresponds to the number of examples in each bin. Each cell contains about 3% of samples under uniform sampling of the stimulus space. | 110 |

| | | |
|------|--|-----|
| 3.17 | Left: Histogram of the orthogonal distances between the chosen teaching examples and the boundary. Right: Histogram of the distances between teaching samples and a line orthogonal to the boundary (corresponding to the irrelevant feature in the RB condition or the opposite diagonal in the II category). The values presented were rescaled to maximum possible distance from the subjective boundary. | 112 |
| 3.18 | Ratio of standard deviations. Numerator: orthogonal distances of teaching examples to the subjective categorization boundary. Denominator: orthogonal distances of teaching examples to the line orthogonal to the subjective categorization boundary. Pedagogical prediction is that the ratio should be smaller than 1. | 112 |
| 3.19 | Proportion of boundaries compatible with the teaching examples. Each dot is a dyad. Crosses mark the sample means. Grey dots correspond to dyads in which one of the participants did not choose a linearly separable example set and were not included in the analysis. | 113 |
| 3.20 | Upper panel: Area of polygons inscribed by the teaching examples, summed over the two categories (excluding overlap). Areas are scaled by the total stimulus space size. Each dot is a dyad. The black cross represents the observed sample means, and the purple cross represents the mean expected from random sampling. Given the structure of the II category, 6 examples can be sufficient to cover the entire stimulus space (for each category). However, the maximum area that can be covered in the RB condition with 6 examples and labels consistent with the true category is 50% of the total area. Lower panel: Distribution of area for random sampling of two triangles from the corresponding categories. Experimental data are overlaid in black (each dot is a participant; vertical jitter added for visualization). | 114 |
| 3.21 | Upper panel: Distance to boundary of the two closest examples labelled assigned to different categories. Measurements were scaled to the unit square. Each dot is a dyad. Black crosses mark the observed sample means, and purple crosses represent the mean expected from random sampling. Lower panel: Distribution of metric under random sampling of three examples in each category. | 115 |
| 3.22 | Upper panel: Relationship between active learning and teaching performance. Active learning is measured by the average distance of the queries to the subjective boundaries across blocks. Lower panel: Relationship between classification accuracy in the final block and teaching performance. Each dot is an active learner. Only teachers who provided linearly separable teaching sets were included. None of the correlations were significant. | 117 |
| 3.23 | Violin plots showing the distance to the subjective boundary as a function of example order. Distances have been scaled to the 0-1 range for every participant. Vertical lines mark medians. Active and yoked learners are pooled together within a category structure. | 119 |
| 3.24 | Inferred participants priors over the possible boundary locations. | 129 |
| 3.25 | Expected information gain from the second active learning query against the maximal information gain given the first query. Each dot is a query selection. Colors denote conditions. | 130 |

| | | |
|------|---|-----|
| 3.26 | Chance level for entropy reduction during teaching as a function of boundary location (blue line) against observed performance(gray line) and performance broken down by group, with prior active learning in red and without prior active learning in black. | 131 |
| 3.27 | Estimated r against proportion of decisions compatible with the labelled data (left) and against normalized proportion of decisions compatible with the labelled data (right). Each cross represents a participant and conditions are denoted by colors. | 132 |
| 3.28 | Within dyad performance differences. Each cross represents a participant. . . . | 133 |
| 3.29 | Estimated r against proportion of decisions compatible with the labelled data (left) and against normalized proportion of decisions compatible with the labelled data (right). Each cross represents a participant and conditions are denoted by colors. | 133 |
| 3.30 | Teaching information gain as a function of the r parameter. The fitted OLS regression line for each condition. | 134 |
| 3.31 | Visual illustration of the decision-bound model used to fit the participants' subjective boundaries (shown in purple) from their category choices. | 135 |
| 3.32 | Categorization accuracy across participants and blocks. | 136 |
| 3.33 | Categorization. Each row is a participant. | 140 |
| 3.34 | Active learning. Each row is a participant. | 142 |
| 3.35 | AIC values for a strong sampling response model (fitting just the probability of responding with a certain category) and the decision-boundary model used in the analysis. Each dot is a participant. Black dots represent active learners and grey dots represent yoked learners. | 143 |
| 3.36 | Average query distance to (individual) subjective boundaries. Bars represent standard errors of the mean. The dashed line is the expected query distance under strong sampling for each corresponding category structure. Asterisks are displayed for blocks where the average query distance differs from the dashed line significantly in a one-sample two-tailed t-test. | 143 |
| 3.37 | All examples chosen by participants. Colors denote the categories chosen by the participants before designing the example. | 144 |
| 3.38 | Teaching. Each row is a participant. Area colors mark the correct category. Dots are stimuli offered as examples. The black lines are the true boundaries and purple lines are the subjective boundaries inferred from the last test block. . | 145 |
| 3.39 | Frequency of teaching examples across the stimulus space (pooled across all participants). Color intensity corresponds to the number of examples in each bin. Each bin contains about 3% of samples under uniform sampling. | 146 |
| 4.1 | Example trial from the observation phase. | 174 |
| 4.2 | Mutual information of agents. | 177 |
| 4.3 | Visualization of predictions for collaborator preferences. Absolute values are only illustrative. | 182 |
| 4.4 | Collaborator choices across the conditions | 183 |
| 4.5 | Estimates of the agent's accuracy and performance on the categorization task. . | 186 |
| 4.6 | Violin plots showing the participant's estimates for the agent's accuracy and performance on the categorization task by participant gender. | 187 |

| | | |
|------|---|-----|
| 4.7 | Participants' estimates of the two agents' accuracy in the control conditions (replicating main task conditions E and F). Crosses denote sample means. In the rightmost plot of within participant differences against zero, circles denote means and lines correspond to medians. | 187 |
| 4.8 | Example trial from the observation phase. | 194 |
| 4.9 | Slide summarizing the performance of the two agents so far. These summaries were presented to participants every 30 trials. | 196 |
| 4.10 | Example trial from the Betting phase. Here, the participant chose the advice of the agent represented by the green avatar. | 196 |
| 4.11 | Task design. Test and Control blocks were run within-participant. Conditions A and B were between participant conditions. Numerical differences presented here were scaled for visual presentation in the experiment (50% confidence corresponding to the center of the scale, and 100% to the extremes). | 197 |
| 4.12 | Left: Relationship between confidence and probability of being correct for agents with the same accuracy, but different marginal confidence. Right: Predictions for decisions as a function of confidence differences between advisers and condition. | 199 |
| 4.13 | Fitted curves from each participant's logistic model on decisions. Dashed vertical lines represent predictions for the boundary based on recalibration and bold lines are group averages. | 203 |
| 4.14 | Posteriors distribution for the categorization boundary. Vertical lines represent the MAP and horizontal intervals include the 95% HDI. | 204 |
| 4.15 | R^2 for predicting individual confidence judgments from the confidence of the potential advisers. Each dot is a participant, all conditions are overlaid (red: cond A test; blue: cond B test; gray: cond A control; black: cond B control). | 205 |
| 4.16 | Relationship between confidence of chosen adviser and average confidence of participants. | 206 |
| 4.17 | Figure extracted from Yates et al. (1996): Calibration graphs of the forecast-outcome data used in the experiments | 210 |
| 4.18 | Example of a stimulus set used in Experiment 1 generated using code from Op de Beeck et al. (2001). | 210 |
| 4.19 | Avatars used in Experiment 1 | 211 |
| 4.20 | Origins of different confidence profiles | 211 |
| 4.21 | The effect of gender on collaborator preferences. | 212 |
| 4.22 | AIC comparison used for participant exclusions. | 213 |
| 4.23 | Regression weights for predicting participants' confidence based on the confidence of advisers. Each dot is a participant. | 214 |
| 4.24 | Fitted weights from logistic regression of participant decisions on the two advisers' continuous judgments (signed confidence). Each dot is a participant. | 215 |
| 4.25 | Variance explained(R^2) for models of participants' confidence. Each dot represents a participant. | 216 |
| 4.26 | Participants' judgments as a function of adviser's judgments. Each line corresponds to a participant. | 217 |

| | | |
|------|--|-----|
| 4.27 | Participant estimates for the two advisers' accuracy and confidence. Participants did not see numerical values on the scales, the half slider length was rescaled for plotting to the 50-100 range. Each dot is a participant. Crosses represent sample averages. Vertical and horizontal lines are the ground truth values. | 218 |
| 4.28 | Observation phase of pilot. | 219 |
| 4.29 | Participants' individual boundaries, inferred by regressing binary choices on the difference in adviser confidence. Dashed horizontal lines represent the predictions based on recalibration of confidence. | 220 |

List of Tables

| | | |
|-----|---|-----|
| 3.1 | Dimensions and boundaries used in pilot for Experiment 1 | 127 |
| 4.1 | The confidence and informativeness (mutual information, MI, in bits) of agents across the experimental conditions. The maximal information gain achievable is .88. Larger differences in MI are expected to result in more extreme preferences. | 178 |
| 4.2 | Agent calibration profiles in Experiment 1 for conditions A-D | 180 |
| 4.3 | Agent calibration profiles for Experiment 1 for conditions E-H | 181 |

Chapter 1: Introduction

1.1 Brief overview: How teachers help learners

It has been proposed that humans have an innate sensitivity to pedagogical guidance and that a propensity for teaching others is at the core of human cumulative culture (Csibra & Gergely, 2011; Csibra & Gergely, 2006, 2009). According to this account, pedagogy is both human-specific and universal among human cultures. Whether or not this is the case¹, from a normative standpoint, learning from teachers who are knowledgeable, well intentioned, and attuned to the learner should be vastly more efficient than individual self-guided learning or purely observational social learning.

An immediate reason for this benefit is that good teachers foster good circumstances for learning. Empirical developmental work demonstrated that humans scaffold the individual learning of children (Wood et al., 1976) by modifying the learning environment in a way that supports the young learner and is tailored to their current level of understanding.

More interestingly, beyond scaffolding, the teacher and the learner's recognition of being in a pedagogical context can further facilitate learning. In line with this, even infants are sensitive to which acts are communicative and relevant for them through monitoring of ostensive signals such as pointing or the use of 'motherese' (Csibra, 2010). The first proposal for how pedagogical contexts influence learning is that it constrains, or, in other words, sets a prior on the kinds of things that learners expect to be taught. Specifically, pedagogical cues cause learners to

¹It should be noted that there are multiple definitions and operationalizations of teaching in the literature. We direct readers to Kline (2015) (and multiple subsequent comments) for a useful classification of mentalistic (Tomasello et al., 1993), cultural (Paradise & Rogoff, 2009) and functionalist (Thornton & Raihani, 2008) teaching definitions, as well a discussion of findings in an evolutionary framework.

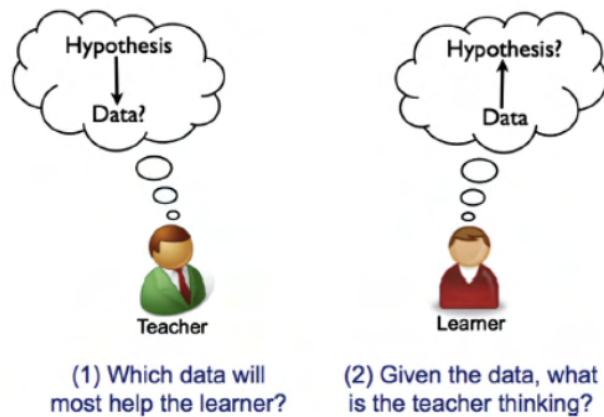


Figure 1.1: Schematic representation of pedagogical reasoning from Shafto et al. (2014).

automatically infer that they are being taught something relevant that is useful and generalizes beyond the current interaction. For instance, in studies of communication about or interaction with a referent object, the prediction is that ostension leads young children to infer that they should learn something about the referent (as opposed to something about the teacher’s mental state in relation to the referent) and that they are taught an essential property about the referent that will generalize to the object class. This allows learners to make interpretations that go beyond observed data (e.g. Futó et al., 2010; Gergely et al., 2002) and learn about latents that map onto “generic and shared knowledge” (Csibra & Gergely, 2011).

The second proposal is that pedagogical contexts influence how teaching actions are produced and interpreted by teacher-learner dyads. Shafto et al. (2014) drew on the rational agent framework (Anderson, 1990) to provide a normative computational-level Bayesian model for teaching. The ‘rational pedagogy’ formalism is very closely related to rational approaches to pragmatic reasoning in communication (Frank & Goodman, 2012)² and similarly presupposes a representation of the mental states of the other member of the dyad (following Tomasello et al., 1993 and unlike Csibra and Gergely, 2009). Under this account, both learners and teachers reason about how pedagogical evidence is generated and about each other’s pedagogical sampling expectations (see an illustration in Figure 1.1). The learner makes inferences assuming that the most informative evidence was presented to them and the teacher generates the

²The only difference is that in the RSA model, the depth of Sinkhorn scaling is 1-2, whereas for the rational pedagogy model, iteration is performed until convergence.

most informative data in light of the learner’s expectation of informativeness (Shafto et al., 2014). Thus, the teaching-learning solution is found through a recursive reasoning process. If a fixed point is found in this iterative process, a pedagogical solution is reached. It should be noted, however, that this is a computational-level account that is agnostic about algorithmic implementations.

Experimental evidence is also accumulating to support the rational pedagogy account, at least in adult-to-adult teaching games, where the behavior of both teachers and learners matched the qualitative predictions of the normative model (Shafto et al., 2014).

From a Bayesian lens, these two lines of research have focused on how pedagogical contexts influence learners’ priors and the likelihood of pedagogical actions. Therefore, the two proposals outlined above are easy to integrate into a common probabilistic framework of pedagogy, and their relative influence remains to be determined empirically.

1.2 Motivation

There are myriad ways in which more knowledgeable and helpful others, teachers, can assist learners in making better inferences - they can demonstrate novel actions, highlight the relevant features that should be tracked by learners for generalization, provide certain examples instead of others, give positive or negative feedback for learner actions. This thesis focused on the human ability to teach by providing (pedagogically sampled) examples. However, the rational pedagogy framework used here is very general as it can formalize cooperative information sharing by any means and apply to a wide range of teaching actions.

Not only is example giving an important means of teaching both in formal and informal pedagogical settings, but it is also an understudied behavior and a complex problem to solve from a normative perspective.

There is a considerable gap in the research devoted to teaching compared to learning and associated inductive biases. Despite the existence of relatively full-fledged normative accounts of what constitutes a good teaching set in the machine learning literature (e.g. P. Wang et al.,

2020) and a long-standing interest in the design of intelligent tutoring systems in education research (e.g. Smith and Sherwood, 1976), we still know little about the behavior of human teachers and how it affects learners. While we can expect humans to be keen teachers, there are potentially meaningful ways in which they will depart from the normative predictions given the complexity of the task and the demands placed on human teachers. The potential deviations from normative accounts will be summarized later in the chapter, and briefly discussed here. Future work may propose teaching models accounting for such deviations within the bounded rationality and computational rationality framework (Gershman et al., 2015), or perhaps it may be necessary to step out of the rational pedagogy framework altogether to characterize human teaching behaviors.

To begin with, little is known about how people explicitly sample from their representations. Explicit sampling for example giving is different from, but can be related to work over the past decade on mental sampling for approximate Bayesian inference making (e.g. Lieder et al., 2012; A. N. Sanborn and Chater, 2016; Vul et al., 2014; J. Zhu et al., 2018) and the explanations it offers for apparently suboptimal behaviors such as probability matching, anchoring, and various reasoning fallacies. It is likely that example generation, much like memory recall or estimation, is subject to consequences of algorithmic implementations of sampling (and the nature of distributions that need to be sampled) such as stochasticity and autocorrelation. Therefore, explicit sampling will result in a combination of unintentional and intentional (pedagogical) modulations, which may subsequently impact the extent to which learners can benefit from pedagogically sampled information.

More importantly, to be effective in their pedagogical sampling, teachers must fulfill a tall order. They need to acquire the true generative model of the task (or a close enough approximation thereof) and decide on the relevant teaching goals since a perfect transmission of information from the teacher to the learner is not feasible in practice. Moreover, teachers need to build a model of the learner and their understanding of the task to tailor teaching to what the learner already knows and how they integrate the evidence they receive in their existing representation. For instance, even in the seemingly simple case of teaching a categorization boundary which

will be used in experiments presented in the thesis, the teacher needs to have a naïve theory of category acquisition which includes assumptions about the prior of the learner over the kind of boundaries and their location.

In a subset of pedagogical situations, both the learner and the teacher independently converge to the correct model of the task. Teaching is then easy (or easier at least) because a model of the learner is not required to be effective - the teacher can substitute it for their own task model. This is a scenario of common ground that is rarely encountered in practice, but was generally the setup of most of the experiments testing the rational pedagogy predictions which relied on highly constrained abstract teaching games with adults. Further, it is clear that the pedagogical sampling model provides a principled way to pinpoint optimal teaching solutions for a range of well-defined problems, and may provide a good account for a range of automatic teaching behaviors in humans. However, it is unclear whether pedagogical sampling underlies on-the-fly ecological teaching or whether in practice teaching is enabled by shared communicative conventions based on general principles independent of the learner's mental state or alternative hypotheses, for instance, matching sufficient statistics of example sets to those of the distribution to be taught.

In what follows, ways to formalize teaching will be described, alongside with relevant experimental evidence and potential challenges.

1.3 Are teachers optimal? Setting the problem

As artificial intelligence is proliferating, it is attempting to solve the sort of complex induction problems that characterize real-world tasks, and thus facing similar challenges as human learners. For instance, while there is now wide access to large image datasets to provide to object classification algorithms, labeled data is costly as it is reliant on tedious human annotation labor so it must be used wisely. This has given rise to the sub-field of machine teaching, which is dedicated to solving the inverse problem of machine learning: finding an optimal learning sequence.

Therefore, there is fertile ground for asking questions about what scaffolds the remarkable success of human learners and teachers, and whether substantial analogies can be drawn to the work in machine teaching that seeks to formalize this problem. Machine learning work in areas such as the explore-exploit dilemma has often informed cognitive theories and behavioural experiments and proposals have even been made for neural mechanisms involved (Cohen et al., 2007). In contrast, while machine teaching has been studied in the machine learning community for quite some time (Goldman & Mathias, 1996; Hegedus, 1995), only recently have there been a number of limited attempts to compare the predictions of these models to human teachers (Khan et al., 2011) or to harness insights from human teaching such as problem decomposition to facilitate the complexity of speed achievable with machine training (Simard et al., 2017).

Coming from a different perspective, the cognitive science literature has been more successfully (cross-)fertilized by emerging computational work in hierarchical Bayesian inference. There are suggestions as to how induction learning could be made tractable and indeed some models nearly reach human-level performance in specific tasks targeting, most notably one-shot learning (Lake et al., 2015), structure discovery (Kemp & Tenenbaum, 2008) and theory of mind (Baker et al., 2017). Stemming from this tradition, Bayesian computational level models have also been used to discuss how induction could be made more effective and efficient by manipulating different aspects of the data which are presented to the learning algorithm. Most pertinent for the current topic, research has focused on the stronger inferences afforded by placing assumptions on the way teachers generate data (Shafto, Goodman, & Frank, 2012).

It is worth considering the differences and commonalities between the classical machine learning literature and the recent Bayesian modeling in cognitive science on how they frame the optimal teaching problem. In the Bayesian inference framework (Shafto et al., 2014), the task of teachers in induction learning problems can be loosely formalized as drawing samples from their internal probabilistic representations such that when observed by the learners, these samples will bring the learners' representations as close as possible in all the relevant dimensions to those of the teachers.

At the core of this formalism is the idea that the most representative set of examples are

those that maximize the posterior probability of the target model and that the work of finding representative data boils down to matching sufficient statistics (Tenenbaum & Griffiths, 2001). Of course, this inevitably couples the teacher's likelihood and the learner's posterior, but they can be solved by fixed-point iteration (sequentially evaluating the likelihood and posterior until convergence) starting from an initial likelihood. ³ Figure 1.1 illustrates the pedagogical process. Shafto and Goodman (2008) start from the assumption that the teacher and learner share the (static) hypothesis space and the prior over hypotheses. The learner acts rationally and changes their beliefs in light of new evidence according to Bayesian updating:

$$p(h|d)_{\text{learner}} \propto p(d|h)_{\text{teacher}} \cdot p(h) \quad (1.1)$$

The rational teacher assumption states that teachers will select examples that would increase the learner's belief in the correct hypothesis (i.e. the teacher's hypothesis). Shafto and Goodman (2008) specify this through soft maximization (the Luce decision rule; Luce (1959)) since choosing only data that maximize $p(h|d)_{\text{learner}}$ is not a robust strategy:

$$p(d|h)_{\text{teacher}} \propto (p(h|d)_{\text{learner}})^\alpha \quad (1.2)$$

In the Shafto model, α is a constant (known to both learner and teacher), rather than a parameter to be estimated. As α increases, the likelihood is sampled according to sharper distributions. In particular, when $\alpha \rightarrow \infty$, the data maximizing the posterior are selected, and when $\alpha \rightarrow 0$, data are selected uniformly from all examples that are consistent with the hypothesis (weak sampling). In this sense, they interpret the α parameter as the greediness of the teacher, with the best examples being chosen for larger values of α . On the other hand, if the teacher is very greedy, this can foster very efficient learning of the hypothesis, but it narrows the space of hypotheses considered by the learner.

It becomes apparent that solutions for optimal teaching depend on learning behavior and vice versa. Thus, in order for an event to be optimal, the learner and teacher equations need to

³Note that multiple solutions are possible

be jointly solved. Intuitively, the solution for this system of equations can be found iteratively. Starting from weak sampling ($\alpha = 0$) or probability matching ($\alpha = 1$), the likelihood of the teacher and the posterior of the learner can be recursively estimated until the estimates become stationary, having iterated to a fixed point.

The setups in which these models are tested generally involve a simple and highly constrained learning problem presented in the form of a non-interactive game because this is the only regime in which they are tractable. However, in essence, they provide a blueprint for how an optimal teacher ought to behave given the specific task constraints and allow for experimental testing of model predictions against human performance.

The machine learning literature has predominantly formalized teaching as a problem of efficient communication⁴. This has been largely applied to concept learning problems where the goal is mapping objects to binary labels in order to identify a target concept given a class of possible target concepts (a concept space). In contrast to the standard approach under which the labels given to learners are drawn independently at random (or independently from the positive extent of the concept), in machine teaching, the evidence is sampled for learners with the goal of minimizing the number of examples needed to perfectly identify a concept. The perfect identification of the target is achieved when a set of examples is consistent with the target concept but inconsistent with all other concepts in the class⁵. The smallest such sample is referred to as the teaching dimension of the concept given the concept class, and the teaching dimension of the class is the maximal teaching dimension of all the concepts included in it.

Intuitively, we can think of the average teaching dimension of a class as quantifying the data size necessary to achieve a given level of effectiveness (Goldman & Kearns, 1995). There are many variants and useful extensions of this to cooperative settings, most notably the recursive teaching dimension which exploits the hierarchical structure of the concept class Zilles et al. (2011). The advantage of the algorithmic teaching approach is the possibility of deriving proofs

⁴This is the most prominent direction of research, although work relevant to teaching has been presented in various areas from model compression, curriculum learning, knowledge distillation, human-machine interaction or adversarial ‘teaching’ in cybersecurity.

⁵Note the direct connection to the weak versus strong sampling in the experimental literature on induction Navarro et al. (2012).

of learnability. On the other hand, tractability is brought about by assumptions of deterministic settings which are of limited interest for applications to behavioural models. Yang, Yu, et al. (2018) more recently proposed an extension towards probabilistic models of learning and an associated index, the Transmission index, that measures the communication effectiveness of a pair of probabilistic inference and data selection process. Further, this index has been applied to cooperative inference where both teacher and learner can exploit the knowledge that the other is cooperative (i.e. the data selection and the inference process are intertwined as described above) resulting in a Cooperative Transmission Index. This index can reveal for what forms of the likelihood it is possible to achieve optimal cooperative transmission, which could be used in the future to what problem sets can be effectively addressed by teaching.

It is encouraging to see converging solutions emerging from the two communities; however, what both these approaches have in common is that they proposals for optimal teaching behavior do not take into consideration the limitations of human teachers and learners. While the optimal strategies seem intuitive and would guarantee marked improvements in learning, they are at best not an easy feat and at worst intractable for human teachers and learners. In what follows, I will elaborate on the difficulties associated with providing a satisfactory solution for optimal teaching considering the constraints of real-life tasks. Furthermore, informed by the constraints bounding rationality, I will review the limited behavioral evidence for the optimality of human teachers.

1.4 Optimal teachers in the real world

1.4.1 Learning about others

If the teacher's goal is to transfer a representation (or key elements of a representation that are crucial to solve a task) to a naive learner by providing carefully selected examples, s/he must also have a model of how the learner incorporates the new evidence and what is their starting representation. Conversely, the learner needs to be aware of the teacher's knowledgeability and intentions, and exploit this when making inferences about the information conveyed. There-

fore, based on the data made available, they also need to make inferences about the intentions of the teachers and assess their knowledgeability in order to avoid being deceived. Epistemic vigilance plays a crucial part in social cooperation as agents need to generalize from the behavior of others to decide whether they are honest and reliable partners. This is a heavy inferential burden since all these tasks must be performed simultaneously, that is, learning about others while learning from them (Landrum, Eaves, et al., 2015) or teaching to them, respectively. A way to mitigate this is by sequentially performing inference about the data when we are certain about teacher's intent and making inferences about the teacher when the information is familiar. However, this is possible only in a very small subset of situations and it eschews the issue of how the initial assessments can be performed (other than via inbuilt biases).

While a full blown theory of mind might indeed not be required for pedagogical learning to be possible, the two clearly have a tangled trajectory (Bass, Bonawitz, et al., 2017). To account for this circularity, simplifying assumptions have been proposed, most notably the rationality assumption (Shafto, Goodman, & Frank, 2012; Shafto et al., 2014). The complication of inferring the other's model of the task and how they incorporate evidence is cast aside as it is assumed that the learner will perform Bayes-optimal inference over predefined common-ground data and hypothesis space. The teacher will present the samples that (soft-)maximize the learner's posterior probability for the hypothesis to be taught. Conversely, the learner will assume that the teacher is well-meaning, informed, and Bayes-rational in producing samples.

The first of the two assumptions, the rational learner hypothesis, has gained support in the developmental literature, as evidence amounts to suggest that even infants possess the building blocks to infer the relative probability of generative models from the observed data (Gopnik & Bonawitz, 2015). Further, even young children are sensitive to sampling assumptions (Gweon et al., 2010; Xu & Tenenbaum, 2007), making it plausible that learners can use statistical or pragmatic information about how the data were generated, such as the presence of a teacher, to decide on the appropriate likelihood function. This could mean that given the same evidence, a learner could reach different inferences when s/he assumes a teacher is present or not. And crucially, making the pedagogical assumption should lead to stronger inferences given a fixed

number of teaching samples. A simple example to illustrate the difference between purposeful pedagogical sampling and random sampling is the rectangle game Shafto and Goodman (2008) where the learner must guess the location of a rectangle on a game board based on being presented with two dots that are included in the rectangle. Teachers were shown the rectangle and asked to choose two examples for a learner. The pedagogical expectation, matched by behavioural data (Shafto & Goodman, 2008), is that teachers provide two opposite corners of the rectangle as positive examples, and that learners infer that the target is the smallest rectangle that includes both of the provided dots if they are told that someone else has chosen the examples (but not if they are told they were randomly chosen).

It should be noted that in the simple pedagogical games used by (Shafto, Goodman, & Frank, 2012; Shafto et al., 2014), the learners had the same ability level as the teacher and could easily put themselves in the teacher's shoes and consider what would be the optimal teaching strategy. However, most teaching occurs in scenarios of imbalance between the ability and knowledge of the teacher and learner, where the optimal strategy is possibly opaque for the learner. On the other hand, it is worth mentioning that even four-five year old children were shown to be able to choose evidence in accordance with a cooperative or competitive goal (Rhodes et al., 2015) so it is possible that the predictions hold even situations of knowledge imbalance.

In situations of common ground violations, there might still be a benefit of pedagogical sampling due to the statistical structure of the examples provided and the inferences they can lead to, but this would be potentially greatly reduced if the sampling assumption is mismatched (i.e. the teacher is sampling from a different type of distribution or over a different set of distributions). One example is the case of the prototype teaching game where a univariate Gaussian distribution describing a new concept is taught by providing a limited number of examples. Based on seeing these positive instances of the concept, the learners need to rate the likelihood of other examples belonging to the concept. If the teacher is constrained to use three examples, the best examples that result of the rational pedagogy formalism are the mean and two (symmetric) extremes showing the range of the concept. If one is looking for a Gaussian

distribution, this is an intuitive strategy, but one could also think that these examples could be optimally chosen to teach a distribution with three modes.

As mentioned above, learners must be epistemically vigilant, but what are the cues to trustworthiness, what is the developmental trajectory of this ability and the computations underlying it? The developmental literature has shown that 3-year-old children prefer familiar teachers even if they provide incorrect evidence (Corriveau & Harris, 2009), perhaps because of a long-term track record of providing accurate information, but this trend is reversed by age four when accuracy trumps familiarity. Shafto, Eaves, et al. (2012) can explain this developmental pattern by extending their previous model by assuming that an unhelpful (cf misleading) teacher would seek to minimize the learner's posterior probability for the hypothesis-to-be transferred. Rather than incorporate a truth heuristic (whereby children assume others are trustworthy), they use familiarity as a prior over the teacher's knowledgeability and helpfulness. Through lesioning the model and eliminating the ability of learners to rely on helpfulness (knowledge-only model), they can account for the modulations in preference based on familiarity and accuracy found empirically. This model has been further tested with adult participants using rating of informants as opposed to relative preference for an informant (Warner et al., 2011), showing that participants could infer intent spontaneously. While this is a simple scenario which teachers being either helpful or not, knowledgeable or not, it makes an important point about the explanatory power of the joint inference framework. Extending this work, Landrum, Cloudy, et al. (2015) show that as children grow older, they start to use the graded quality of the evidence (beyond the dichotomous classification as misleading/ helpful), defined as the typicality of an example and diversity, in their evaluation of teacher credibility. Of course, other cues can be useful for establishing trust in a teacher, such as their associated confidence and how it relates to their accuracy.

Another interesting factor that has been shown to influence the rating of informant helpfulness is whether s/he conveys information in the most effective and efficient way. At the worst data offered by teachers ought to be inductively sufficient, enough to make reliable inference but not more. Shafto, Gweon, Fargen, and Schulz (2012) argue that learners should penalize

teachers who provide exhaustive evidence that only marginally serves to reduce uncertainty, due to the cost of demonstration. They incorporate a prior on the data in the model which is an exponential parametrized by the number of demonstrations and the cost of each individual data point. They fitted this model to participant data in an informant helpfulness rating task, revealing that the estimated parameters corresponded to an learner expectation of sufficient for accurate performance. Learners did not apply a penalty for additional data, but also did not prefer maximal data. It is very likely that the data cost insensitivity stems from the design of the task which does not explicitly penalize for longer response times. Indeed, it is unclear what is the origin of this cost and why it affects learners (as opposed to just teachers) unless it is conceptualized as computational cost of information transfer. In line with this, when children are asked to teach another person, they modulate the amount of information they provide depending on the knowledge level of the learner (Gweon, Shafto, et al., 2014) or their goals (Gweon, Chu, et al., 2014) by omitting unnecessary information, and irrelevant information respectively. Further, they act according to a naive utility calculus, whereby they try to maximize the utility of the learner by balancing the costs and benefits of a teaching plan. In their experiment, this translates into preferentially teaching how to manipulate a toy the learner finds more interesting and that would be more difficult for the learner to figure out on their own.

There are other ways in which teachers could come to learn what their students believe that rely on different assumptions. Rafferty et al. (2015) propose a inverse reinforcement learning framework, whereby teachers observe learners' actions as they try to achieve a goal transparent to the teacher from a given starting state. By formalizing action planning as a partially observable Markov decision process, the teacher can compute a posterior probability distribution over the learners transition model. In other words, teachers will gain knowledge about the learner's subjective beliefs about how their actions will alter the state of the world. Such a model can then be used by teachers to guide feedback and correct potential mispecifications of the transition model. In the context of a simple game (learning how to fly an alien spaceship using button presses), the authors show that their model can infer learner beliefs nearly as accurately as humans. Further, automated feedback informed by the model (feedback that targets the but-

ton about whose function the learners were most likely to be wrong) lead to better performance than uninformative feedback (random feedback that does not take into account the learners' current beliefs). However, Rafferty et al. (2015) did not then also compare the feedback offered by human teachers with that produced based on the model. Clearly, this goes beyond teacher models that only monitor learner success or failure, and do not consider the sequence of actions leading to that result. On the other hand, it suffers from the same limitations as the Shafto, Goodman, and Frank (2012) proposal: the action, state, hypothesis space as well as the goal are shared by teacher and learner and the knowledge state of the learner is stable. If this later constraint is to be removed, a probabilistic learning model would be needed. Lastly, it is important to note the differences between this approach that proposed by Shafto and collaborators. Here, the teacher makes inferences solely through observation and the learner's performance is enhanced due to the targeted feedback provided, orthogonal to what the learner believes about the teacher's knowledgeability or helpfulness.

1.4.2 Representation of the task environment and hypothesis generation

While the Shafto, Goodman, and Frank (2012) proposal for optimal teaching is expressed at the computational level, there are objections that can be brought about from the algorithmic level. Namely, another feature of optimal teaching (e.g. the pedagogical sampling account) is that inference is performed over a pre-established hypothesis and data space which is common ground for teacher and learner, they also share a prior, and the inference must be performed on all hypotheses and datapoints simultaneously.

However, a crucial and nontrivial aspect of learning is the generation of appropriate hypotheses which is influenced by the representation of the task environment. Moreover, it is unlikely all hypotheses are considered simultaneously. A widely accepted proposal is that learners engage in a stochastic search through a space constructed by applying random modifications to the originally entertained hypothesis when it no longer fits the observed data Goodman et al. (2011) and Markant and Gureckis (2014). The learner then needs to check if this new hypothesis is consistent with the currently available data.

Indeed, part and parcel of the teaching process is guiding the search for a suitable hypothesis that may not even have been considered initially. Even when they have a superior representation of the task environment, teachers additionally need the ability to monitor the hypotheses currently entertained by learners and have a grasp of the dynamics of sequential hypothesis generation.

1.4.3 Teaching is extended in time

Teaching is a process that extends over time, with examples presented sequentially leading learners to update their representations gradually. Moreover, teaching goals themselves are not immediate, requiring planning to reach a long-term objective that can be served by a wider repertoire of teaching actions than just providing data. To complicate matters further, this long-term objective is generally vague - we prize teaching that enables students to generalize their knowledge to novel problems that we cannot fully anticipate.

Traditionally, both in machine learning and cognitive science, learners have been thought of as batch inference makers with little attention being paid to the dependence of samples or their order. In fact, using the same pedagogical assumptions as in batch presentation, but updating the learner's posterior after each sample and solving for this new chained posterior and likelihood after each subsequent sample leads to completely different predictions as well as strong serial dependencies.

Sequentiality is starting to be addressed in interactive machine teaching. For instance, using naturalistic egocentric video of infants playing with toys in a word learning experiment (Yurovsky et al., 2013), an omniscient interactive machine teacher with sequential data presentation (Liu et al., 2017; Yurovsky et al., 2013) managed to generate a similar temporal structure of the frames as observed naturally (e.g. continuous bouts of the same object instance), and unsurprisingly the learners converge much faster than when the (standard stochastic gradient descent) learner receives random inputs.

1.4.4 Learners are active

Learners are not passive receptacles of knowledge; they interact with their teachers (and their environment) and direct the learning process by requesting more information or specific types of evidence that they believe is most diagnostic. The benefits of learner-centered exploration have long been extolled (Gureckis & Markant, 2012), and we know that exploration is reduced after instruction (E. Bonawitz et al., 2011a) due to the expectation that the teacher has conveyed all relevant information.

In light of this trade-off, are there circumstances should we do away with teachers and let learners explore? Yang and Shafto (2017b) addressed this issue in a computational analysis which pitted optimal active learning (here, choosing a query to maximize expected information gain) against optimal teaching across a range of scenarios. When the teacher and student are aligned (the learner bias is known, rational inference is performed over a fixed hypothesis and data space), teaching will be at least as good as active learning. When the concepts are ambiguous (overlapping), teaching maintains an advantage. Further, similar performance was achieved with exploring and teaching under learner misconception, meaning that the learners concept space does not match that of the teacher, which is also normatively correct. When the teacher was misaligned, teaching lead to poorer performance. In essence the take home message is that teaching is likely going to benefit learners (or at least not harm them) unless there is a fair degree of misalignment. However, one should keep in mind that the relative effectiveness of these methods will depend on the concept size and complexity.

1.4.5 Redundancies can be optimal for bounded learners

While we have emphasized the need for plausibly bounded teachers, the reverse of the coin should be considered. Namely, that optimal teachers ought to take into account that even resource rational learners are bounded cognitively. Learners could be bounded in terms of memory capacity for the retrieval of exemplars during decision-making, noisy perception, and compressed representation. For instance, while you might intuitively think that in a categori-

sation task you should present labeled examples as close as possible to the boundary, however if the learner is limited in its capacity to retrieve exemplars at test this could lead to higher erroneous classification (Patil et al., 2014).

It can be that teaching strategies which appear to be subpar with respect to the informativeness principle alone (i.e. give only the minimally sufficient information) are more useful in practice to learners. In this sense, the problem of efficient teaching mirrors that of efficient communication, where there is a larger body of work departing from traditional Gricean pragmatics, to recast, for example, referent identification tasks, as collaborative tasks (e.g. Rubio-Fernandez, 2019) in which redundancies actually facilitate discriminability.

Another example is curriculum learning. For example, in a simple one-dimensional binary classification problem, an interesting effect is that it is sometimes more helpful to forgo the optimal teaching dimension strategy (offering only examples close to the boundary) and start by giving examples at the end of the continuum and then move to increasingly harder to classify examples. This method received experimental support in phoneme discrimination (McCandliss et al., 2002) where second-language speakers could learn much better after initial overarticulation.

Interestingly, a recent study by Khan et al. (2011) in which participants were asked to teach a one dimensional classification task based on a continuous subjective feature (their own rating of the ‘graspability’ of objects and chosen boundary location) to a humanoid robot showed that curriculum teaching is among the preferred teaching strategies. In fact, none of the teachers used the optimal boundary strategy even when explicitly instructed to use as few examples as possible and under no time pressure to make the decision. The authors suggest that in fact curriculum learning is a good heuristic in optimization to avoid poor local optima (Bengio et al., 2009). An idea that needs to be explored further is that the curriculum learning is optimal if the learning progress is to be maximized (i.e. expected error is minimized) after every iteration. This corresponds to a “win stay, lose shift” (Gibbs) strategy where a new hypothesis is chosen randomly from consistent hypotheses when the current one is no longer consistent with the teaching set. Given the psychological plausibility of this strategy, a direct experimental test of

their predictions would be particularly informative.

Related to this, for higher dimensional/ hierarchical concepts, it is meaningful to consider not just how the examples presented are useful, but also how the (cooperative) omission of information shapes inference (Searcy & Shafto, 2016). This can aid in deriving an intensional concept from an extensional definition which scales down the complexity of teaching (specifically, it scales according to the complexity of the concept and not the concept space). Generalization to the narrowest level that includes all examples (Xu & Tenenbaum, 2007) and sensitivity to sample diversity in learning and teaching (Rhodes et al., 2008; Xu & Tenenbaum, 2007) support this account. However, partially inconsistent results have also been presented. In a teaching task (Cakmak & Thomaz, 2010) using concepts defined as conjunctions of compound tangram feature values in which some features are relevant and others not, teachers failed to achieve exact object identification (although as mentioned in a previous section, learners exhibit a preference for such teachers). This means that participants stopped teaching before sufficiently pruning the concept space to only one consistent concept (the target). While teachers introduced variability in the irrelevant features, they did so less than expected based on optimality. The rate of concept pruning was also suboptimal as teachers used about 3-5 times more examples than needed. Further, even when the optimal strategies (start with positive example, vary as many irrelevant features as possible in a positive example, vary relevant features one at a time) were described to the participants, they failed to teach efficiently.

Moreover, while the strong assumption thus far has been that teachers are maximally helpful and well-meaning, it is worth relaxing this assumption by considering the gains of the learner and the effort incurred by the teaching action. A starting point is that effort ought to be proportional to the number of examples a teacher needs to provide, although realistically it is going to depend on the teaching goal and the problem to be solved. An interesting development in that direction generalizes the teaching dimension to probabilistic, noisy learners that have potentially infinite hypothesis spaces (Zhu, 2013). Crucially, the loss function is arbitrary and effort, as a function of the teaching dataset, is added to the loss that needs to be minimized. They solve the optimization problem (over teaching examples) for Bayesian learner models in

the exponential family and exemplify teaching one dimensional classifiers, a Gaussian mean, multinomial distributions, multivariate Gaussians, and model selection.

1.5 Conclusions and further directions

We reviewed a small and concentrated experimental literature on teaching. At the moment the evidence in favour of the optimality of teachers relies on test scenarios that are over simplified and abstract, but there is a promising future in this line of inquiry. A first issue to clarify though is whether the results of Khan et al. (2011) and Cakmak and Thomaz (2010), with more complex stimuli and ecological settings, are replicable and can be reconciled with the teaching games conducted by Shafto et al. (2014). This will be addressed in Chapter 2.

Different possible reasons for departures from the predictions of rational agent models were enumerated, primarily the need for teachers to solve additional problems such as: making inferences about the social partner, the lack of common ground, need for long term planning and for balancing the relative costs of teaching. There are also likely limitations at the algorithmic level as not all hypotheses will be considered or evaluated simultaneously.

We welcome interest from computational fields in formalizing more realistic, resource-rational standards for optimal teaching that exploit inferential assumptions of learners and teachers. Perhaps one of the most important contributions of the computational approaches is to fully specify all the interacting conditions that affect (or need to be met for) teaching to be successful and teasing them apart: the (different) model, representation, starting conditions, learning algorithm, loss function corresponding to the teacher and learner, plus the goal and intent of the teacher, available action space for both actors, and ability to track the learner's behaviour and/or knowledge state. While this may seem overwhelming, there are now some sufficiently sophisticated proposals for what optimal behavior should be given these constraints (e.g. Zhu, 2013) and so there is a window of opportunity for experimental work that tests the predictions of these models.

The pertinent question then becomes how these departures will affect learners in practice?

The education literature offers some clues to that as teaching behaviors were generally analysed not in relation to normative predictions, but in terms of concrete effects on actual learners. The findings supported the fact that real-world formal teaching is good, but far from ideal. Teachers were shown to tailor their behavior based on the structure of the problem to be taught (e.g. increasing the level of difficulty), but to rarely seek to understand the subjective model of the learner (e.g. by asking how they arrive at their incorrect solutions) and rather provide feedback based solely on whether the learner is objectively correct (Chi et al., 2004). In line with this, human teachers were found to be about as effective as simple natural-language-based computer tutoring VanLehn et al. (2007). Against all expectations, an extensive metareview failed to find a meaningful advantage of human tutors in STEM (teaching one student synchronously) over intelligent tutoring systems (Kurt, 2011) in terms of learning gains. These automated systems provided feedback to students on their correctness, divided the problems to be solved in preset substeps and provided a printed explanation for each step if the student's response contained errors - by no means reaching the level of sophistication implied by formal teaching accounts proposed here. There was also no evidence of a difference between expert and novice human tutors, or an effect of constraining the interactivity between tutors and learners. This is good news for students given the recent proliferation of online teaching, but it underscores the potential limitations of human teachers in complex tasks. These results may also speak to differences between teaching behaviors in formal settings and teaching in informal settings (cooperative learning).

In the following experimental work, we will focus on ways in which human teachers may potentially alleviate the burden of this complex task, specifically the teacher's challenge of building an adequate model for the learner (Chapter 3) and the additional cues that learners may use to make decisions about which teachers to rely on (Chapter 4).

This also should inform questions about whether it is possible to teach the teachers to compensate for their limitations. The answer seems to be trivially yes since we already spend considerable resources training educators in formal settings, but in the only two experiments presented in this review where this was attempted, the results were not encouraging.

Chapter 2: Sampling for teaching and learning

2.1 Introduction

Exceedingly little is known about how humans explicitly sample information from their representations with the intent to teach, despite the fact that a lot of teaching taking place through the offering of examples and demonstrations.

The Introduction outlined both a normative model for how teachers can explicitly and purposefully produce examples (Shafto & Goodman, 2008), concerns that the assumption of rationality does not do sufficient work to address all challenges of identifying good teaching sets, as well as mixed experimental evidence on the optimality of human teachers. A straightforward first step taken in Study 1 was to replicate key previous findings with the aim of clarifying disagreeing findings.

The corresponding and equally challenging problem facing learners is adequately incorporating the complex process that generated the data received from teachers in their inference making. Learners must form expectations for pedagogical sampling based on the communicative context. Additionally, learners need to find ways to cope with the unintended features of samples generated by another person.

There is evidence that humans are sensitive to differences in the sampling processes involved where evidence is gathered from other humans as opposed to the environment. Children and adults in Xu and Tenenbaum (2007) generalized differently as a function of whether they

acquired information from pedagogical compared to random sampling. Meng Yuan and Fei (2017) have shown that perceptions of agency are positively correlated to perceptions of non-randomness in binary sequences. Furthermore, Gweon et al. (2010) have shown that children as young as 15 months can distinguish weak from strong sampling.

Here, we focused on one feature of pedagogically sampled sequential data and, generally, of samples produced by humans (Gilden et al., 1995; Lieder et al., 2012) and many Bayesian samplers, but which is also ubiquitous in the environment - autocorrelation. Study 2 tested whether humans optimally estimated the means of sequentially presented series of stimuli as a function of the underlying generating process: i.i.d. or strong temporal autocorrelation.

2.2 Study 1: Teachers

We replicated two tasks that involved pedagogical example giving to teach prototype concepts (Experiment 2 of Shafto et al., 2014) and rule-based concepts (Khan et al., 2011). These simple tasks were chosen since they both have theoretically optimal teaching solutions and have also been studied with human participants, with different conclusions.

Shafto et al. (2014) found that human teachers were able to successfully convey both two-dimensional rule-based concepts and univariate prototype concepts. In the rectangle game, participants could choose two labeled examples (where the label was inside or outside the rectangle) to convey the location of a rectangle in a larger grid. Teachers predominantly chose the optimal solution, closely marking the opposite corners of the rectangle. In the prototype game, participants acting as teachers were presented with a series of one-dimensional stimuli drawn from a Gaussian distribution that they were told defined a novel concept and they had to choose a batch of three of stimuli to teach the concept to an imagined learner. Here, the optimal solution is to mark the central tendency of the distribution alongside two extreme symmetric stimuli¹.

However, Khan et al. (2011) have instead found human teachers used various suboptimal

¹This is optimal when the hypothesis to be taught is at the center of the hypothesis space.

strategies and greatly overestimated the number of examples needed to teach a rule-based concept. In the Khan et al. (2011) experiment, participants were asked to teach a one dimensional binary classification task based on a continuous subjective feature (the participants' own rating of the 'graspability' of objects) to a humanoid robot (see Figure 2.3). This is perhaps the simplest task on which human teaching has been tested, and yet it produced surprising results. Conceptually, this teaching problem is equivalent to teaching the ends of a line segment, a simpler task than teaching the location of a rectangle, so the results are at odds with previous findings of Shafto et al. (2014).

It is unclear what the source of this discrepancy is. First, it could be that the more ecological setting of the Khan et al. (2011) task, teaching a robot with no common ground about a real (though abstract) concept may be different from engaging in a reasoning game about an imaginary concept. Related to this, it is possible that when attempting to perform pedagogical sampling of multivariate objects, participants intentionally or unintentionally considered irrelevant features of these objects. For instance, participants may have wanted to highlight features that make an item ungraspable (e.g. being voluminous, heavy, or difficult/dangerous to approach).

Second, results could also be an artifact of participants aiming to teach the underlying ranking of stimuli on the graspable-ungraspable continuum rather than a binary classification. This could be an explanation for the use of a linear strategy according to which participants labeled many or most of the stimuli, moving left-to-right or right-to-left through the stimulus array.

On the other hand, some of the participants in Khan et al. (2011) were clearly engaging in a curriculum teaching strategy. That is, they started by providing extreme examples on either side of the stimulus space and then gradually offered examples closer to the boundary. While this strategy is suboptimal for an optimal learner observing the data in batch mode, it is optimal if the teaching objective is to minimize expected error at every time step (with no planning). Further, as mentioned in the introduction, curriculum learning is widely observed in human teaching and is sometimes necessary to achieve learning.

Lastly, all teaching games showing optimal teaching decisions have constrained the example set size cardinality to the minimally sufficient number. It is possible that when constrained to a given number of examples, participants would be prompted to make the optimal choice on the one-dimensional teaching task as well.

Therefore, in order to reconcile the findings in the literature, we set out to replicate the experiment of Khan et al. (2011) in a more controlled setting with a computerized format, using several categorization dimensions besides graspability. Furthermore, to assess whether constraining the example set size is the crucial factor behind inconsistencies in the literature, participants were asked to generate teaching examples both in an unconstrained manner (as many as they sought necessary) and constrained to two examples.

2.2.1 Replication of Shafto et al. (2014): Prototype teaching

Methods

Participants

Twenty participants (aged 20-26 years old, 14 female), all Hungarian native speakers who were fluent in English, were recruited through a local student organization (MADS) and paid for their participation (1200HUF). Subsequently, all participants participated in an unrelated experiment on statistical learning. Given the size of the observed effect, a smaller sample size than in the original experiment ($n = 28$) was deemed sufficient.

Procedure

Participants were first shown a set of 27 stimuli: heavy black lines of various lengths drawn inside a thin black circle (instead of a rectangle as in the original experiment), on a white background. Stimuli were individually printed on paper cards that were randomly placed in a 3x9 grid on a table. The positioning of the cards differed for each participant.

Written instructions (identical to the original experiment) were translated for the partici-

pants in Hungarian (and were also available in English):

"In this experiment, you will see a random assortment of "widgets"—objects consisting of a circle with a line inside it. The circle is always the same size and the line always starts at the same point, but the line varies in length. If you look closely, you can probably see that widgets are more likely to have lines of some lengths than others. Imagine that you had to teach somebody about the distribution of line lengths that one sees on widgets, but could only do so by showing them three of the widgets in front of you. Which three widgets would you choose?" (Shafto et al. (2014), p. 16).

The experimenter summarized the instructions verbally (in English) and checked comprehension. Following this, participants chose a set of 3 cards to show as examples to an (imagined) future participant in the experiment by putting them in a specially marked area of the table. Participants were allowed to change their mind about their card choices as many times as they wanted.

On average, the entire task took approximately 10 minutes.

Stimuli

The 27 line lengths were sampled (in Matlab R2012b) from a Gaussian distribution with a mean of 14.5 mm and a standard deviation of 3.5 mm.

Each card, 90 mm wide and 50 mm high, contained a black line stimulus (1.5mm thick), enclosed in a circle with a radius of 31 mm. The lines started 1 mm from the left-hand side of the circle and were vertically centered.

Data Analysis

To quantitatively test whether data were best fit by strong sampling or pedagogical sampling, the probability for triplet exemplars to be generated from the two models was calculated as described below, and a likelihood ratio test was applied to determine if the observed distribution of triplets was more likely to be generated from the multinomial distribution resulting

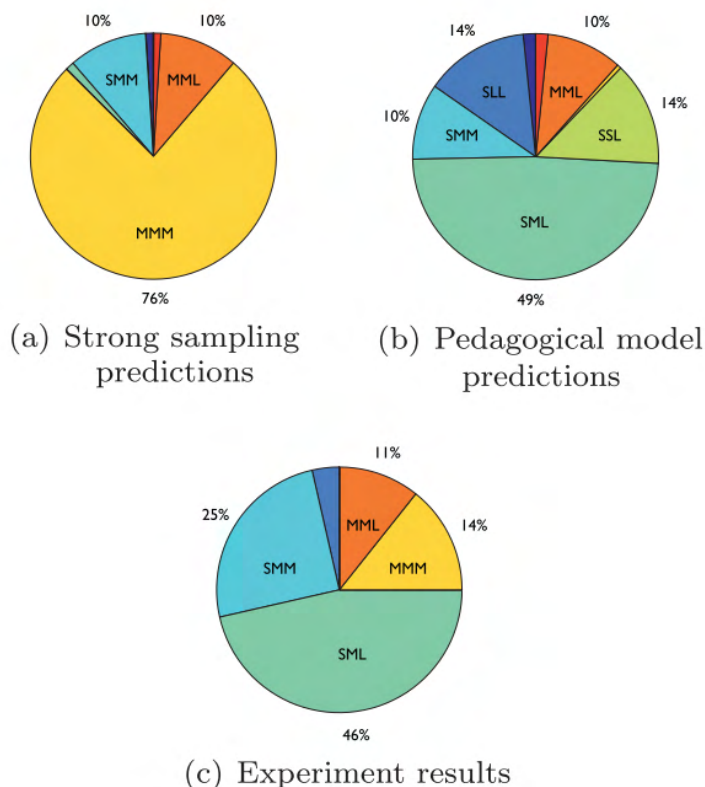


Figure 2.1: Empirical results of Shafto et al. (2014) (Fig. 7) along with model predictions.

from pedagogical or strong sampling.

In both sampling implementations, the possible teaching example sets consisted of all possible order-invariant triplets out of the 27 stimuli (in total 3,654 sets).²

Under strong sampling, which also provided the starting likelihood for the data for the pedagogical model, the probability of each triplet set under the true sample parameters was calculated by factorization assuming that example choices were independent.

The hypothesis space for pedagogical reasoning consisted in all the distributions defined by the mean and standard deviation of these triplet sets. Null standard deviations were set to .01 and other standard deviations were halved³. A uniform prior was used and the likelihood was a discretized Gaussian (probabilities of the 9 values were renormalized to sum to 1). Ini-

²In the original experiment model implementation, a smaller model of the task was used which started with only 9 possible examples, evenly spaced through the stimulus range. Examples were binned into small (smallest 2), medium (middle 5), and large (largest 2) to compute triplet probabilities. Predictions were closely matched.

³The standard deviations were halved in order to obtain values in a similar range to the variance of the hypothesis to-be-taught. Without halving the SD, the probability of choosing SML triplets decreases, but qualitative differences between pedagogical and strong sampling remain.

Replication results: Triplets chosen by teachers

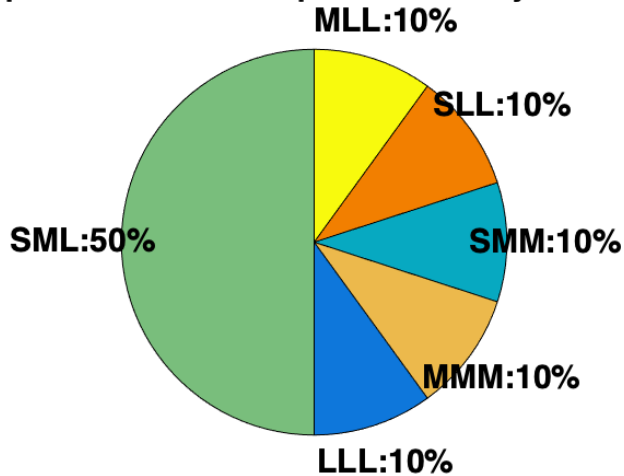


Figure 2.2: Triplets of examples chosen by the teachers. L: large, M: medium, S: small.

tial triplet likelihoods were calculated, for every hypothesis, by factorization assuming that the choices were independent. The α parameter was set to 1. Lastly, examples were binned in the same way as for the behavioral analysis into small (smallest 6), medium (middle 15), and large (largest 6), and probabilities were summed to compute probabilities for the ten possible triplets (e.g. SML, SMM, SSS). Resulting likelihoods for hypotheses with means within .5mm distance from the stimulus mean and standard deviation within .25mm from the experiment value were pulled together.

Results

The chosen examples were classified as small (smallest 6), medium (middle 15) or large (largest 6). [Figure 2.2](#) shows the distribution of the example choices. Similar to the original results (see [Figure 2.1](#)), roughly half of the participants chose triplets composed of one small, one medium, and one large example. This is qualitatively consistent with previous results, as well as with the predictions of the rational pedagogy model.

Triples chosen by participants were more likely to come from pedagogical sampling than strong sampling, Bayes Factor (BF_{alt}) = $2.55e+14$, $\chi^2(1) = 66.34$, $p < .001$.

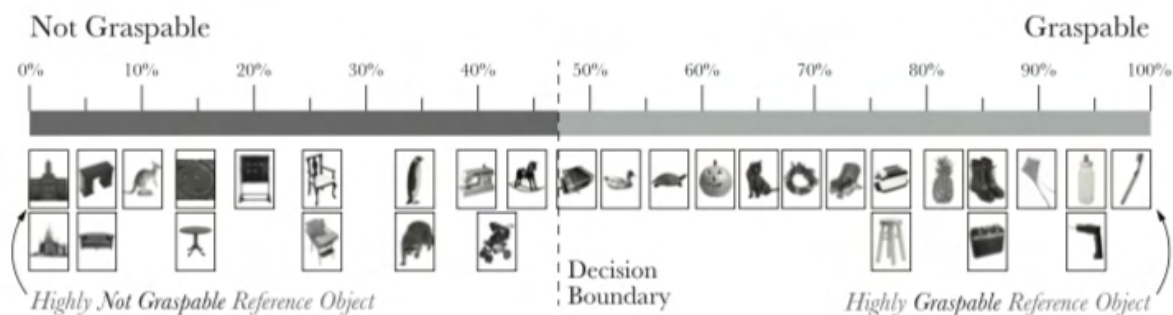


Figure 2.3: Example stimulus array from the Khan et al. (2011) teaching task. Participants ranked images as a function of ‘graspability’ and then chose a boundary between graspable and non-graspable items. The image was reproduced from Khan et al. (2011).

2.2.2 Replication of Khan et al. (2011): Boundary teaching

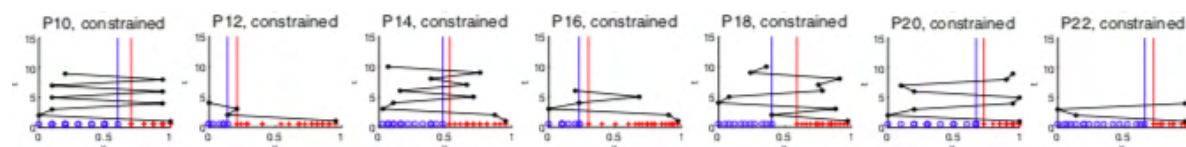


Figure 2.4: The trajectory of the examples chosen by participants conforming to the curriculum learning strategy. The image was reproduced from Khan et al. (2011).

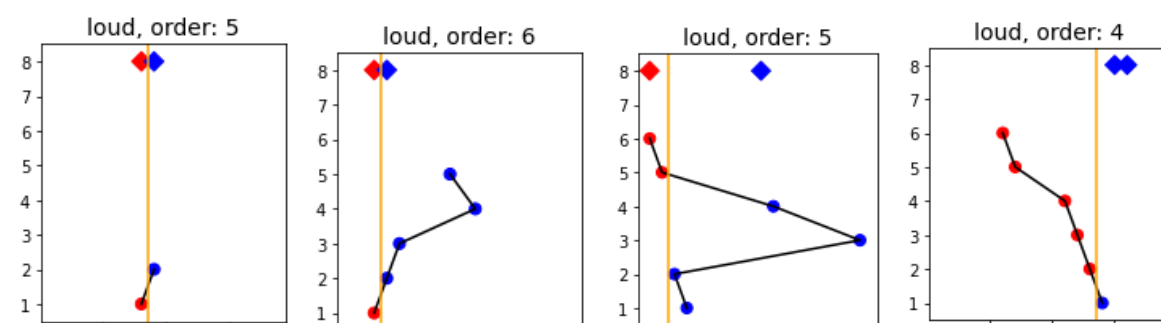


Figure 2.5: The trajectory of the examples chosen by participants. Circles represented example choices in the Unconstrained condition (the order of choice is down-up). Diamonds represent the examples selected in the Constrained condition.

Methods

Participants

Válasszon minél kevesebb példát, csak annyit, amennyi Ön szerint feltétlenül szükséges ahhoz, hogy megtanítsa a másik embernek, hogy mi hangos és mi nem.

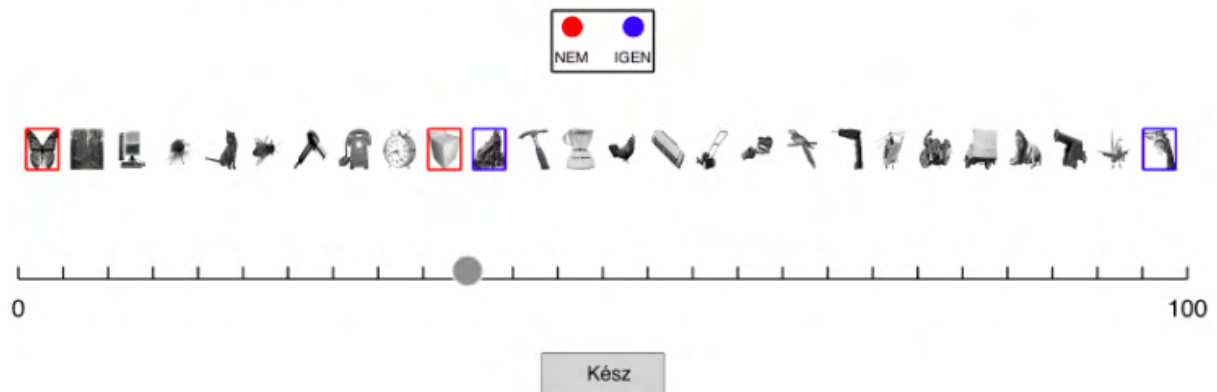


Figure 2.6: Example image array from the teaching task. In this trial, images were sorted according to loudness. The gray dot marks the boundary chosen by the participant. In order to select examples for teaching, the participant had to drag either the red (not loud) or blue (loud) circles across an image, which will automatically mark it as an example by drawing a square of the corresponding color around it.

Twenty five native Hungarian speakers, also fluent in English, aged 18-25 years old, were recruited through a Hungarian student employment association (MADS) and were paid for their participation (1200 HUF/hour). One participant was excluded from the experiment because they did not comply with instructions (choosing the first two stimuli in all trials).

Design

A computerized, Hungarian language version of Khan et al. (2011) was implemented in PsychoPy2. In addition to graspability, five additional dimensions were used to categorize the objects in the images: price (for demonstration), speed, loudness, numerosity, color, and size. The last three dimensions were added to test whether (suboptimal) choices were previously induced by the multivariate and/or abstract nature of the images, the identity of which was irrelevant to the categorization. Half of these dimensions relate to concrete, perceptual quantities, whereas the other half refers to abstract classification rules and realistic, multidimensional objects.

The images were presented in random order in a horizontal line spanning the entire length of the screen, and participants first ranked the items by dragging them into slots in the desired order. The participants then decided where they wanted to place the boundary between the

two categories before teaching. A slider marking all the possible boundaries was drawn on the screen to visually display the potential boundaries.

There were two parts to the experiment. In the first, participants were asked to teach the boundary they saw on the screen to another participant using only as few examples as they thought necessary in order for the other participant to know its location (Unconstrained Condition). In the second part of the study, the trials (and boundaries) were presented once more, and participants were asked to choose only two examples to teach the location of the boundary (Constrained Condition).

Procedure

The experiment started with a demonstration trial in order to familiarize participants with the experimental setup. The first test trial classified the items according to graspability. The presentation order of the graspability trial was fixed in order to avoid any interference from the other trials for the purpose of the replication. The order of the following trials, asking for judgments based on speed, loudness, and numerosity, size, and color was randomized.

In each trial, 26 images⁴ were presented on a 27-inch screen, and were ranked by participants from left to right according to the key dimension. The participants then chose the boundary, which remained visible on the screen until participants finished selecting the teaching examples by dragging category labels to the images. A screen shot of an example selection trial is presented in Figure 2.6.

The task was unspeeded and lasted around 30 minutes in total.

Materials

All images used in the demonstration, graspability, speed and loudness trials were selected from the same dataset of black and white photos of objects (Salmon et al., 2010) used in the original study. Overlap between images used across the trials was avoided as much as possible, with only a handful of images repeating at most once.

The stimuli for the numerosity trial were constructed by drawing varying numbers of small red circles displayed within a larger gray circle (positioned uniformly at random). The stimuli

⁴The original study had 27 images.

were equidistant on the log-linear scale. The corresponding cover story was that the boundary was the a given number of bacteria in a Petri dish required to meet a infection diagnostic. In the color trial, colored circles ranging from blue to red, implemented through a perceptually uniform matplotlib colormap, and participants were asked to find the boundary between what they thought was blue and red. In the size trial, the stimuli were vertical black lines uniformly increasing in length, and the boundary was the size of a battery.

Results

In the Unconstrained condition, participants chose nearly twice the number of examples needed to teach the boundary, although considerably fewer than in the original study, $M = 3.84, SD = 1.95$, see [Figure 2.7a](#). Only 7 participants (out of 24, 29.17%) managed to find the optimal solution with the minimal example set cardinality in at least one trial in the Unconstrained Condition. In the Constrained condition, the number of participants increased to 11, 45.83%. Overall, teachers provided example sets that were consistent with a unique boundary in approximately 57% of the trials.

[Figure 2.7c](#) shows the remaining uncertainty after teaching in the two conditions, that is, the number of boundaries that are consistent with the teaching set provided. The few trials ($n = 16$) in which teachers chose labels inconsistent with the boundary were removed. Uncertainty was lower on average in the Unconstrained Condition compared to the Constrained Condition, as the mean of possible boundaries compatible with the labeled teaching set was 2.06 relative to 5.23. This was also the case within participant (and within trial), as can be seen in [Figure 2.7b](#). There were no stark differences between trials as a function of order in the task or classification dimension. The number of examples chosen in the Unconstrained condition not significantly relate to the remaining uncertainty around the boundary, $\tau = -.08, p = .31$.

Participants did not exhibit the clear qualitative patterns observed originally, that is, clear cut instances of the curriculum or linear teaching strategies. Some sample trajectories of example selections are presented in [Figure 2.5](#).

In order to quantitatively test predictions, the remaining number of compatible boundaries

in every trial was regressed to the manipulated within-participant variables: Condition (Constrained/Unconstrained)⁵, Trial type (Abstract/Perceptual), and the order of the trial. Given the structure of the experiment, a mixed effects model on the trial-level data was deemed most appropriate, with participants as the random intercept and the above mentioned predictors as fixed effects. Further, due to the large proportion of trials with a unique compatible boundary, a zero-inflated Poisson model was used. The model was fitted with the GLMMadaptive library in R, using a log link for the non-zero part and the logit link for the zero part. The random intercept structure improved model fit, $\text{LRT } \chi^2(2) = 38.89, p < 0.001$.

First, whether the choice of examples was constrained or not, had a significant effect both on the zero and non-zero processes: $\beta_{\text{unconstrained}} = 1.51, z = 3.52, p < .001$, and $\beta_{\text{unconstrained}} = -.47, z = -5.61, p < .001$, respectively. This mirrored the descriptive results, namely, the lower number of compatible boundaries for the Unconstrained Condition.

The type of trial, perceptual or abstract, had no significant effect for either component, $\beta = .67, z = 1.63, p = .10$, and $\beta = -.01, z = -.02, p = .98$.

The order effect was significant only for the non-zero component, $\beta = -0.05, z = -2.08, p = 0.04$, suggesting that the number of compatible boundaries decreased with increasing trial order, as expected.

2.2.3 Conclusion

The original results of the prototype game Shafto et al. (2014) have been successfully replicated (in a different culture, but also with young university students), as half of the participants exhibited the optimal teaching pattern. This corresponds to the proportion expected if participants are drawing samples by probability matching from the pedagogical distribution.

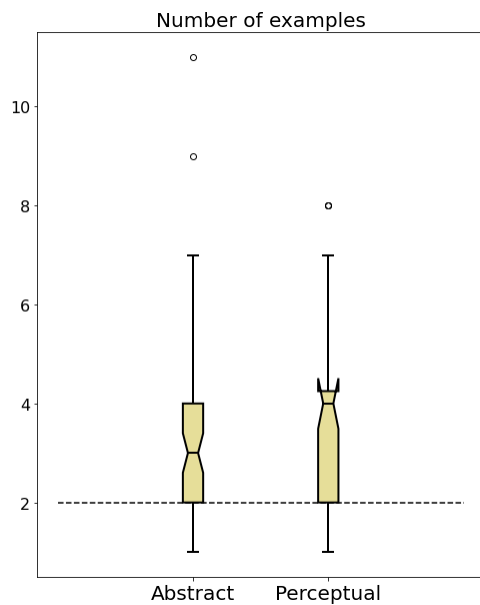
The results of Khan et al. (2011) were broadly replicated, but some significant departures were also observed. We also found considerable inefficiency and diversity in teaching patterns, such as continuing to label data after the boundary was uniquely identified. However, in general, our participants used fewer examples than in the original study (4 vs. 8) and we did not

⁵The number of examples was not included as a covariate given its relation this the manipulation.

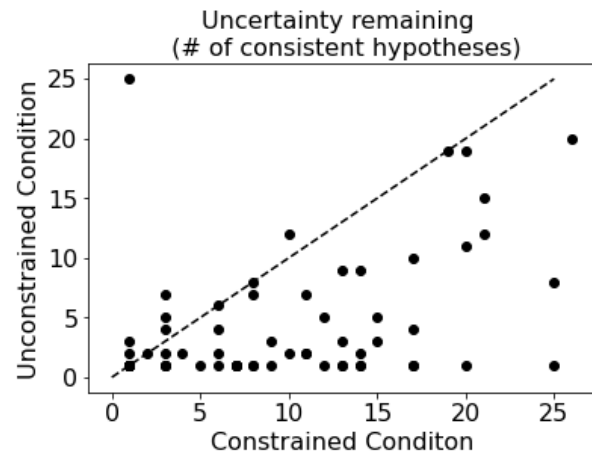
find clear patterns of curriculum or linear teaching. Moreover, while they chose fewer examples, our participants did not maintain the same level of effectiveness. Indeed, one benefit of exhaustive sampling strategies such as curriculum or linear teaching is that the effectiveness of teaching is guaranteed.

The crucial manipulation revealed the opposite pattern to the one expected, as participants reduced less uncertainty about the boundary location when the example set cardinality was fixed as can be seen in Figure 2.7c. Even when prompted with the minimally sufficient teaching set cardinality, participants failed to reach the optimal solution. This suggests that original findings were not due simply to lack of awareness or insensitivity to the costs of teaching (the only cost was spending more time in the lab), but rather participants failed to grasp what would be most helpful to a learner.

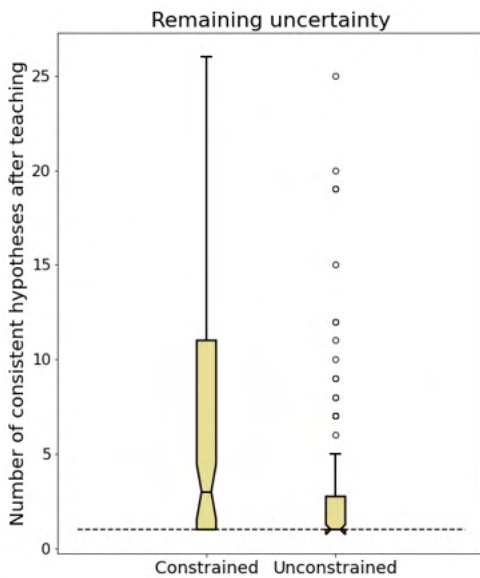
Lastly, our manipulation of the complexity of the stimuli that served as the basis for the classification task did not produce significant differences. This was surprising and underscores the fact that inefficiencies occur even when eliminating uncertainty about what makes a stimulus belong to one category or another, or how it ranks relative to other category members.



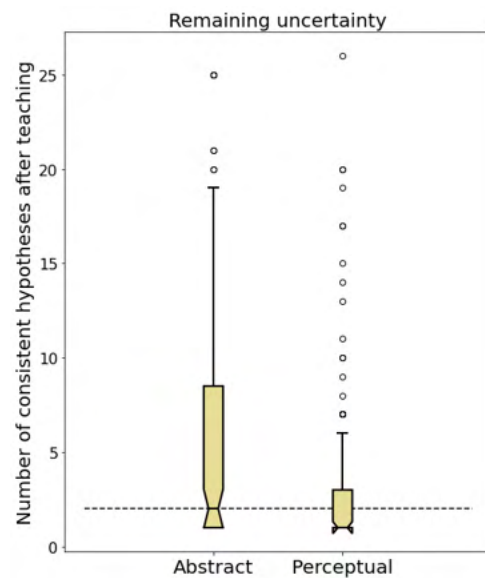
(a) Number of examples chosen in the Unconstrained condition.



(b) Remaining uncertainty as a function of condition. Each dot represents one trial from one participant.



(c) Uncertainty remaining after teaching by condition.



(d) Uncertainty remaining after teaching by trial type.

Figure 2.7: Teaching performance in replication of Khan et al. (2011).

2.3 Study 2: Learners

Serial order effects widely permeate cognitive tasks, from the primacy and recency effects observed in memory recall (Ebbinghaus, 1850-1909) to anchoring in decision making tasks (Tversky & Kahneman, 1974). The cognitive processes and underlying mechanisms driving these effects are as varied as their instantiations. Here, we focus on a potentially novel source of serial effects, evident in estimation tasks which require participants to provide an estimate of a summary statistic describing a sequentially presented stream of data or of a generative parameter governing it. Further, we propose a computational rationale for why such a serial effect leads to efficient performance and inquire whether it can offer a unifying explanation for the preponderance of overreliance on early and late information. In estimation tasks, the optimal strategy for weighting samples at different serial positions depends greatly on the statistical structure of the input. Crucially, when inputs are serially dependent, the optimal weighting pattern exhibits symmetric primacy (overweighting of early information) and recency (overweighting of recent evidence) effects, which increase in magnitude with the degree of dependence between samples. Given the fact that most if not all sequential samples from the natural environment will contain certain degree of dependence, there is a strong motivation for flexible adaptation to the temporal statistical structure. Although recency effects can be easily obtained with optimal models that take into account a changing environment, primacy effects are otherwise difficult to justify on the basis of optimal models.

Lastly, if such serial effects are indeed present in estimation tasks, then it is interesting to inquire whether these effects are related to or orthogonal to the primacy and recency effects observed in memory recall. For instance, it might be possible to modulate memory recall by changing the underlying statistical structure of the inputs.

2.3.1 Overview of experiments

The goal of the following series of experiments was to assess whether, in the context of an estimation task with sequentially presented data, humans are sensitive to the statistical structure

of the inputs they observe and adjust their behavior accordingly. To test this, the strength of temporal dependence in the input was manipulated experimentally and the participants' reliance on samples at different serial positions was estimated.

Moreover, if indeed the observed serial dependence effects are modulated by the covariance structure of the inputs, we expect the effect to follow the direction predicted by the optimal strategy. In the case of estimation tasks which require participants to provide a summary statistic, here the mean, optimal performance is defined in terms of producing an unbiased estimator with the minimum variance. For a given covariance structure of the inputs, it is possible to theoretically derive the weights needed to achieve optimal performance as shown in the section below. The covariance structure used in all the experiments is the first-order autoregressive (AR(1)) which dictates that samples closer in time are more correlated than samples farther in time. Thus, while a clear oversimplification of the sorts of statistical structures observed in the visual environment, this structure encapsulates the key feature of temporal correlations. For the AR(1), as the amount of dependence increases, optimality dictates that there should be a symmetric increase in the relative weighting of the first and last samples of the series. Thus, the general experimental prediction is that when presented with strong dependence between samples, participants will overweight early and late samples relative to the other samples. It is unclear what to expect in the situation when the presented samples are independent. It is possible that in the absence of corrective feedback, and given the ubiquity of correlations in the surrounding environment, which should create a strong prior for dependence of observations, participants will not show the optimal uniform weighting pattern.

In what follows I will briefly outline the experiments conducted with a more detailed description following below. In all experiments, the stimuli within in a sequence, corresponding to a trial, differed in their location on a computer screen which was determined by the AR(1) process as described above. The participants' task was to provide an estimate of the center of mass of each sequence.

In Experiment 1, dots were presented on the screen sequentially, at various locations, and participants estimated the center of mass of the series. Given the heterogenous response pat-

terns, we repeated the experiment with simpler, univariate stimuli, and longer presentation times, manipulating only the dot position along one axis. The variability in the weighting patterns raised the question of whether it is possible to relate the use of optimal weighting strategies with having a better (more accurate/ precise) estimate of the dependence structure in the inputs or where more likely to use that information in order tasks such as prediction. Therefore, a follow-up tested the estimation task alongside a prediction task, which was similar to the estimation task in all respects except the response that was required of participants. The results of the prediction task would provide an additional (perhaps more direct) measure of whether participants were using the correlation structure and at the same time test whether weighting patterns for the estimation task were consistent with prediction performance. However, behavioral results were indicative of over-estimation of correlation in the prediction task evidenced by complete reliance on the last sample and this did not translate into the estimation task.

Experiment 2 presented data generated in the exact same way as in the previous experiment, but using stimuli which had an identity dimension orthogonal from the task at hand, prefacing an experiment investigating memory and estimation links for which a notion of stimulus identity is needed. Again, large heterogeneity was observed in the weighting patterns used by participants. In the follow-up memory task, participants were presented the same stimuli (same location, serial position and shape identity) in separate estimation and memory recall tasks conducted a couple of days apart.

Before delving into the details of the experiments, I address some preliminary theoretical concerns about the validation of the data generation procedure, how to definite (near-) optimal performance and how to estimate the benefit conferred by the optimal strategy.

2.3.2 Defining optimal performance

Given a timeseries X generated from a first-order stationary autoregressive process, such that each element i is computed as:

$$X_i = c + \gamma \cdot X_{i-1} + \varepsilon_i, \quad (2.1)$$

where $\varepsilon \stackrel{iid}{\sim} N(0, \sigma_{innovation}^2)$ and c is a constant, γ is the correlation coefficient. Its resulting auto-covariance matrix is C :

$$C = \sigma^2 \cdot \begin{bmatrix} 1 & \gamma & \gamma^2 & \dots & \gamma^{k-1} \\ \gamma & 1 & \gamma & \dots & \gamma^{k-2} \\ \gamma^2 & \gamma & 1 & \dots & \gamma^{k-3} \\ . & . & . & . & . \\ . & . & . & . & . \\ . & . & . & . & . \\ \gamma^{k-1} & \gamma^{k-2} & . & . & 1 \end{bmatrix}$$

The minimum variance estimator for the mean of the series is given by:

$$\hat{\mu} = (W^T C^{-1} W)^{-1} (W^T C^{-1} X), \quad (2.2)$$

where W is the design matrix for the weights and C^{-1} is the inverse of the covariance matrix. [Figure 2.8](#) shows the resulting optimal weights.

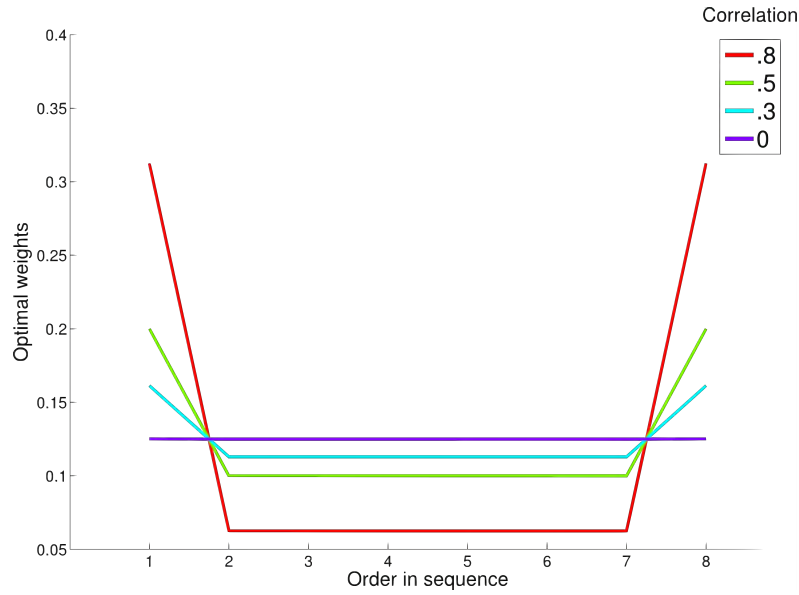


Figure 2.8: Optimal weighting for a series of length ten as a function of the γ parameter which quantifies dependence.

2.3.3 Validation of data generation

The first check performed prior to conducting the experiments is that the parameters used to generate the sequences, in this case the constant c , the correlation parameter γ , and the standard deviation of the error term $\sigma_{innovation}$, can be retrieved from the data samples. This is required to demonstrate that it would be at least theoretically possible for the participants, if they behaved optimally, to get an unbiased estimate of the strength of dependence given the limited time series length and the number of trials used in the experiment. 100 samples were generated using the same parameter values as in the experiment and the same number of trials (300 trials and series of length 8). For each sample, the posteriors over the estimates were calculated using all trials simultaneously and assuming their independence conditional on the parameters. First, the estimation was performed assuming that the generative process is known, that is knowing that each trial utilizes the covariance structure of the AR(1) process. This means that it was not necessary to estimate the entire covariance matrix since it can be generated from the standard deviation of the error term and the correlation coefficient. Since this might be unrealistic, a full estimation of the covariance matrix releasing this assumption was performed with very similar results. The Bayesian estimation was performed in STAN (Team, 2016) with MCMC sampling

(NUTS), using vague priors were chosen for all the parameters, namely a uniform over the $[-1,1]$ interval for the correlation coefficient γ , and a vague inverse gamma (with shape = rate = .001) for the variance of the error term $\sigma_{innovation}$. The posteriors of the parameters based on ten randomly chosen simulations are plotted in [Figure 2.9](#) for the parameters of interest together with the true generative parameters. The posteriors are narrowly distributed around the true value, indicating that it is possible to retrieve an unbiased estimate of the generative parameters. Of course, this does not mean that the participants will be able to retrieve the exact value of the parameters, particularly since they need to jointly infer the statistical structure and the values of the parameters and surely this will take a fair amount of trials that will be used to sequentially update estimates, rather than use all the trials in batch mode. Moreover, each stage of this process can be noisy.

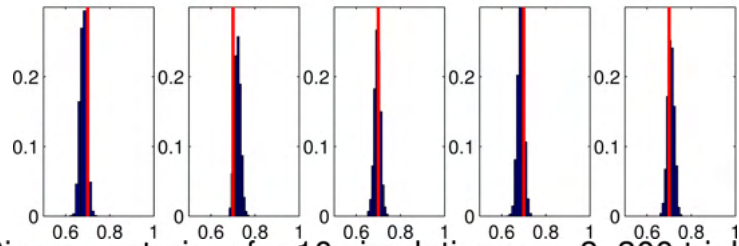
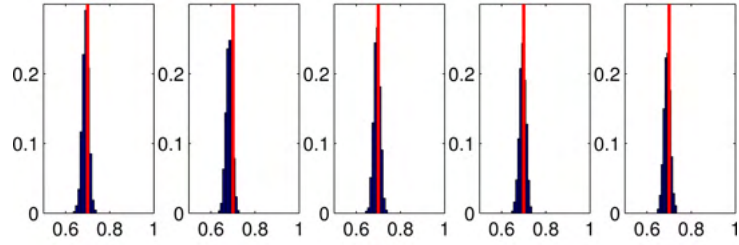
2.3.4 Defining near-optimal performance

Given a given estimate for the correlation coefficient, it is possible to derive the theoretically optimal weights. However, the estimate inferred by the participant may be biased as well as uncertain due to suboptimality in estimation and/or the differences between the statistics of the sample and the true parameter values (variability due to the sampling process). Therefore, it might be unfair to compare behavioral performance directly with the theoretical optimum which assumes participants have access to the true value of the correlation coefficient as manipulated by the experimenter. While addressing the concern of participant suboptimality in estimation strategy is difficult in the absence of a model of the task, it is perhaps possible to address the second concern. Namely, a more appropriate benchmark for optimality could be one that uses the exact covariance matrix of the data sample observed by the participants. Again, this has the caveat of assuming all trials are treated as independent (given the generative parameters) and used simultaneously to inform the estimation. [Figure 2.10](#) presents such ‘empirically optimal weights’ for the simulated datasets with different numbers of trials. As expected, the average of these weights over multiple samples will match the theoretical optimal weights, but there is quite a lot of variability between samples. Moreover, as the number of trials used increases, the

estimated empirical weights are closer to the theoretical weights.

An alternative way to think about this is to use the credible interval around the estimate of the correlation coefficient inferred from the data (assuming the data structure is un-/known), and based on those values calculate an interval containing the likely optimal weights.

Posterior for Rho, Sequence of length 8, 300 trials



Sigma posteriors for 10 simulations, $n=8$, 300 trials

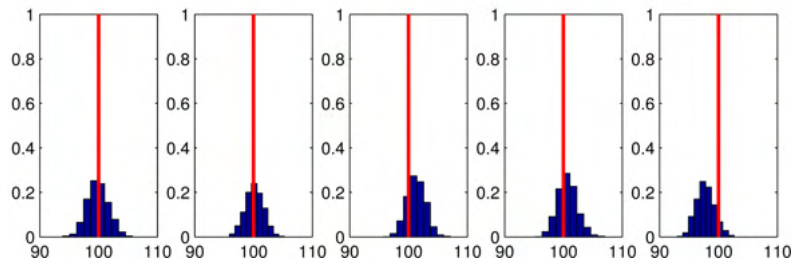
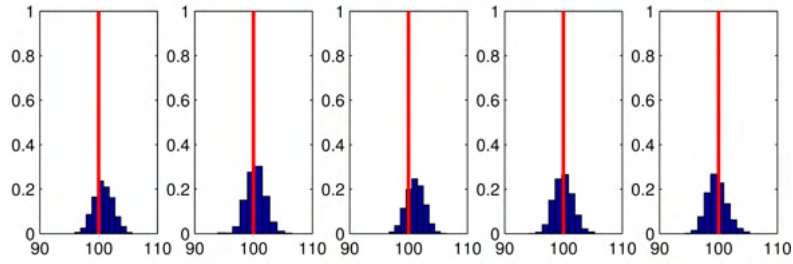


Figure 2.9: The histograms of posterior samples over the correlation coefficient (upper) or standard deviation of the innovation (lower). Each subplot corresponds to one stimulated dataset of 300 trials, where each trial is a series of length 8. The red vertical lines correspond to the true value of the parameter.

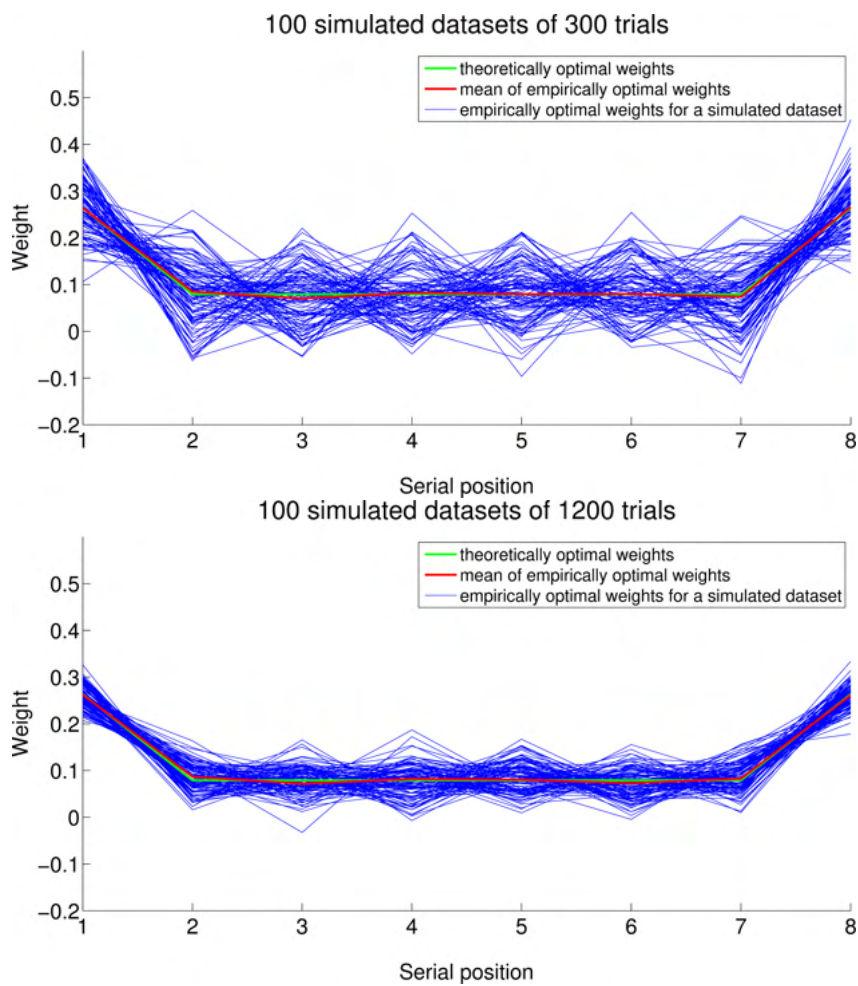


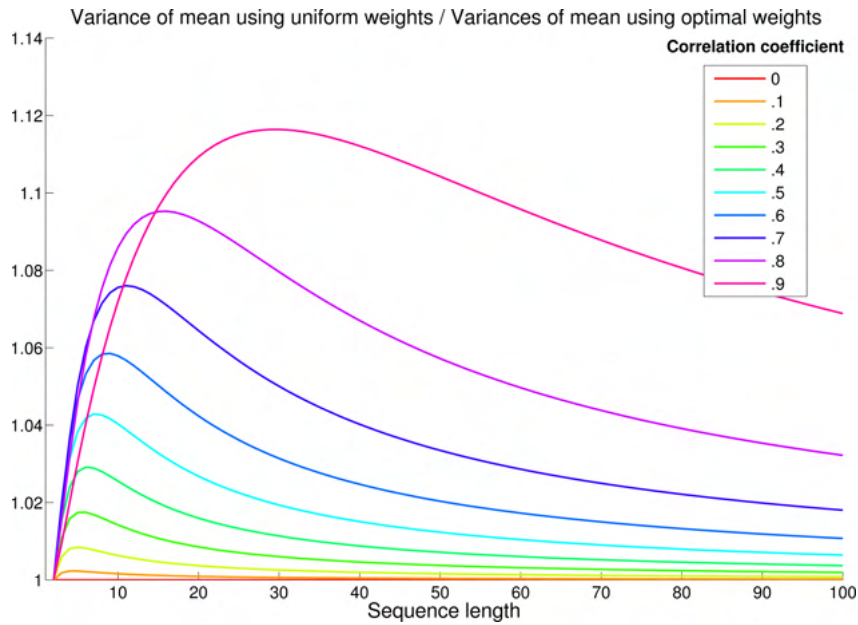
Figure 2.10: Optimal weights derived using the observed covariance structure ('Empirically optimal' weights) of 1000 simulated datasets with sequences of length 8 and using 300 trials (upper) and 1200 trials (lower).

2.3.5 Performance improvement conferred by the optimal strategy

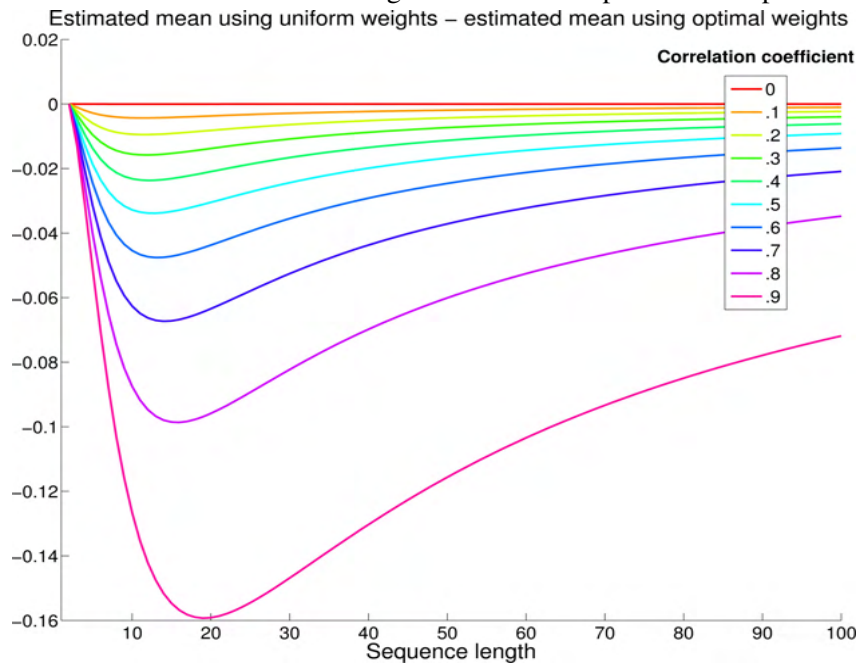
The extent to which humans will comply with an optimal strategy depends on the benefits it confers relative to a competing suboptimal strategy, which may potentially incur a lower cost. While here the notion of a cost is not applicable since all the experiments involve unsupervised learning, it is still possible to make an informed guess about the expected effect based on the relative improvement that could be offered by the optimal strategy relative to a particular suboptimal one. For instance, given a particular value for the correlation coefficient, it is possible to calculate the variance in the optimal mean estimator which uses non-uniform weights and compare it compared to that of the estimator that uses uniform weights. In other words, this quantifies the relative efficiency of using a strategy adapted to the actual dependence present in the data compared to using a strategy suitable for independent data. Additionally, it is possible to calculate the difference in the expected mean when using the optimal compared to the uniform weighting as a function of the correlation coefficient and the strength of dependence.

The ratio of the variances of the two estimators are plotted below ([Figure 2.11a](#)) as a function of sample size and strength of correlation present in the data. As expected, the more correlated the samples, the more advantageous it is to use the appropriate weighting. As expected, this advantage disappears asymptotically, as the number of samples increases.

The bias emerging from using a different weighting strategy than the one used to generate the data, overlooking the presence of sequential dependencies is illustrated in [Figure 2.11b](#). The mean is underestimated by the uniform weighting strategy in proportion to the strength of the dependence.



(a) Ratio of variance of the mean assuming the uniform compared to the optimal weighting.



(b) Bias due to using a suboptimal weighting strategy. This assumed the process constant is 1. The actual difference depends on c multiplicatively.

Figure 2.11: Comparison of uniform and optimal weighting strategies.

2.3.6 Experiment 1

Methods

Participants

42 participants, all Hungarian university students with normal or corrected-to-normal vision, aged between 19 – 26 years⁶, were recruited through a Hungarian student association (MADS) and took part in the experiment in exchange for monetary compensation (rate of 1,200 HUF/hour).

The planned sample size was 40 participants per condition, and was conservatively based on simulation results such that t-tests conducted for every serial position would have adequate power to compare weights corresponding to noisy near-optimal performance to uniform weights⁷. Three participants did not return to finish the experiment on the second day and were replaced.

Ethical approval for the experiment was granted by the Ethics Committee for Hungarian Psychological Research.

Design

Participants judged the center of mass for rapidly presented sequences of dots. Across two within-participant conditions (320 trials), which were completed by participants on different days in counterbalanced order, the generative model for the location of the dots was manipulated.

Each trial consisted in the rapid sequential presentation of 10 dots at various locations on the screen. Grey widely-spaced out vertical and horizontal grid lines were presented on the black background at all times in order to provide a better sense of spatial location, while the dots were white and $.32^\circ$ in diameter. Each trial started with the brief presentation of fixation point (250ms). Each dot was presented for 150ms before it disappeared and was followed by a 50ms inter-stimulus-interval (ISI). Following the last dot in the sequence, there was a 500ms

⁶The gender of the participants was not recorded.

⁷The number of trials was also chosen based on simulation.

pause until the participant could respond. At that point a cross marking the cursor appeared in the center of the screen and participants could move it freely and unspeeded to mark their estimate of the mean position of the series. Once they lifted the pen from the trackpad, the position was considered final and recorded. After the response, there was 1500ms pause until the next trial started.

In one condition, dot locations within a trial were independent and in the other condition, they were strongly dependent. To generate the dot locations, independently for each trial, the centers of mass of the sequence were drawn uniformly from a square uniform probability distribution centered on the middle of the screen (and encompassing 70% of the vertical length of the screen). In the independent condition, 10 dots were then selected from an uncorrelated bivariate distribution with a standard deviation of 2.2 cm, which was kept constant across all trials. In the correlation condition, the dots' positions in vertical dimension were sampled from a univariate Gaussian distribution with the sampled position as the mean and the same standard deviation as above. However, the positions on the horizontal dimension were generated by taking the sampled x position as the starting value of an AR(1) series with $\gamma = .7$. The standard deviation for the error in the AR(1) process was adjusted so that the resulting standard deviation of the samples matched that used to generate the samples in the independent condition. This just involved using a higher standard deviation in the process since the variance of the stationary process is given by $\sigma_{innovation} / (1 - \gamma^2)$. If a generated position was within three standard deviations of the edge of the screen, a new position would be generated until the requirement was satisfied.

Procedure

The experiment was carried out over two consecutive days, with each condition presented on a separate day in counterbalanced order. On the first day, after giving written consent, the session started with a trackpad familiarization task to ensure that the participants were comfortable with the experimental setup and were sufficiently precise in pointing to a target location on the screen using the trackpad pen.

A short demonstration of the task (10 trials) was provided each day which was meant to

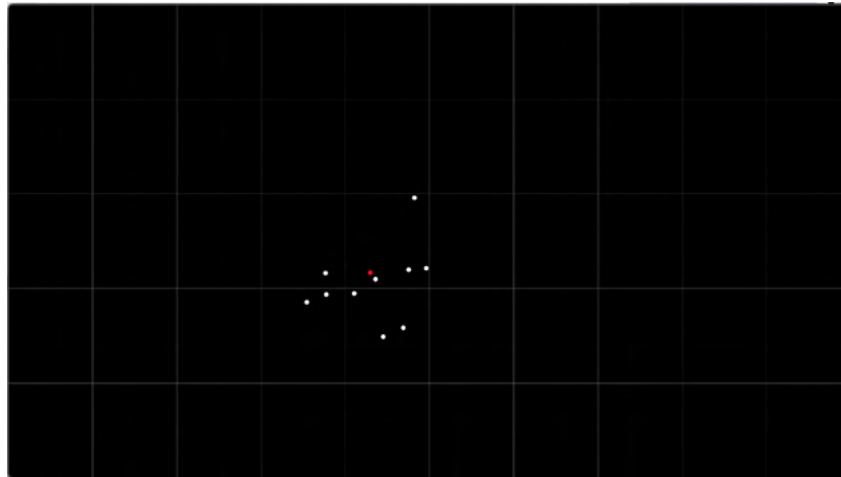


Figure 2.12: Sample end frame from the experiment demonstration.

illustrate the concept of the center of mass as no cover story was used in this experiment. The demonstration data were generated in the same way as the test data, so that the appropriate condition demonstration preceded the test phase. However, the presentation speed was slower to allow the participants to get acquainted with the task.

Within a trial, the ten dots appeared on the screen sequentially but remained visible throughout the trial. After the sequence was complete, the true center of the sample (not the sampling center) was presented on the screen as red dot and the entire scene remained on the screen for 300ms (see [Figure 2.12](#)).

Throughout the experiment, the participants were seated about 50 cm in front of an LCD monitor in a dark room. All in all, the experimental sessions lasted 55-60 minutes, including a break half-way through the experiment.

This experiment and all subsequent variations were coded with the Psychophysics Toolbox 3 extension (Brainard, [1997](#)) for Matlab R2015a. The experiment code and data are now archived but can be available upon request.

Data analysis

Given that our manipulation only affected the horizontal location of the dots, we analyzed the vertical and horizontal estimation separately. For each participant, the distance between the true dot locations and a reference were used as the independent variables on which the distance between the estimate and the same reference was regressed using least squares (no constraints

were enforced).

Given the possibility of aggregation effects, we tested whether every participant's behavior matched predictions. We fit several regression models to the data under different assumptions: symmetric overweighting of the first and last samples, asymmetric overweighting of the first and last samples, overweighting only of the first/ last sample, and uniform weighting. The negative log-likelihoods of these models were then compared for each participant in order to choose the best fitting constrained model for a given participant. We used the Akaike Information Criterion and the Bayesian Information Criterion to choose, for each participant, the most parsimonious model.

Results

As a first sanity check, to confirm that participants were reasonably accurate in their estimation, estimation error was calculated for every participant and compared to a trial-shuffled to ensure that participants were performing the task as instructed (see Appendix [Figure 2.24](#)). For the dependent condition, the estimated means were compared with the expected estimates based on uniform weights, the optimal weighting strategy, and the ground truth ([Figure 2.25](#)).

Further, we report very high levels of variance explained of the unconstrained model, mean 93.25%, for both conditions (see [Figure 2.26](#)). The intercept was not significant for any of the participants, showing no location biases.

Aggregate weights are presented in [Figure 2.13](#) and generally qualitatively match the predictions. As expected, the vertical location estimate was not affected by the experimental manipulation. Furthermore, across all independent conditions, primacy was observed as in previous work.

For the manipulated horizontal location, in the independent condition, the mean weight assigned to the first sample was .134, significantly different from uniform expectation (.1), $t(39) = 3.56$, $p = .001$, $BF_{alt} = 30.35$. The second sample was also overweighted, .126, $t(39) = 3.78$, $p < .001$, $BF_{alt} = 53.04$, but none of the other weights differed from the uniform ($p > .05$).

Primacy was also observed in the the dependent condition, the average weight for the first

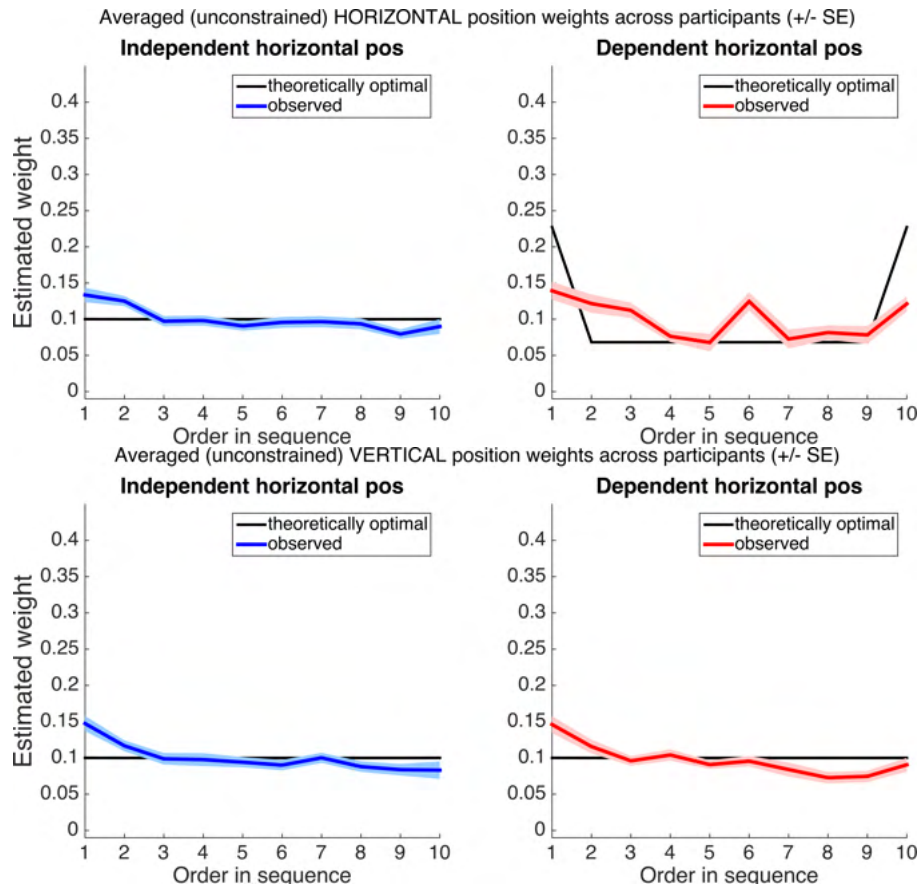


Figure 2.13: Average weights.

sample was .14, $t(40) = 3.39, p = .002, BF_{alt} = 20.19$. Additionally, we observed a small recency effect, .121, $t(40) = 1.99, p = .05, BF_{alt} = 1.01$. Average results conformed qualitatively to the expected pattern of lower weights for samples that were not in the beginning or end of the sequence (4th, 5th, 6th, 7th samples differed significantly from .1). However, there was an unexpected overweighting of the sixth sample, .123, but this was not significantly different from .1, $t(40) = 1.88, p = .07, BF_{alt} = 1.01$.

Within-participant, there was no statistical difference in the weight of the first sample, $\text{diff} = .005, t(38) = .41, p = .69, BF_{null} = 5.38$. On the other hand, weights for the last sample were higher in the correlation condition, $\text{diff} = .031, t(38) = 2.92, p = .006, BF_{alt} = 6.46$.

However, there was considerable inter-individual variability in the relative weights. This was true in both conditions, as illustrated in Figure 2.14 and Figure 2.15. Weight profiles were sometimes similar, but that was not generally true.

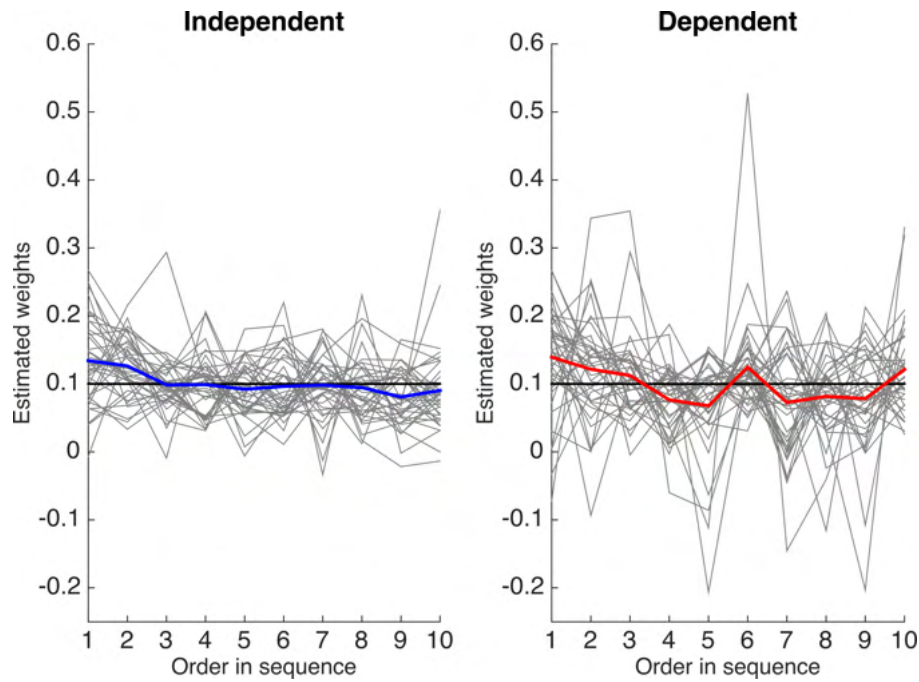


Figure 2.14: Individual weights for the horizontal (manipulated) location alongside the sample average.

In the independent condition, 33 out of 39 participants (~85%) were better fit by the uniform weights model, while only 5 were better fit by the primacy model and 1 by the recency model. Since the primacy model is conservative as it assumes only the first weights is higher than the rest, we also fitted a model in which the weights changed linearly or quadratically. In the AIC comparison, 8 participants were better fit by the linear model (7 with decreasing weights and 1 with increasing weights, from the remaining participants 27 better fit by the uniform model and 4 by the strict primacy).

In the dependent condition, 15 participants were fit by the uniform model, 13 by a primacy only model, 2 by the recency only model, 2 by the asymmetric primacy and recency model, 7 by the linear weights model. As in the independent condition, no participants were better fit by the quadratic weights model. AIC values for all participants are presented in Appendix [Figure 2.28](#).

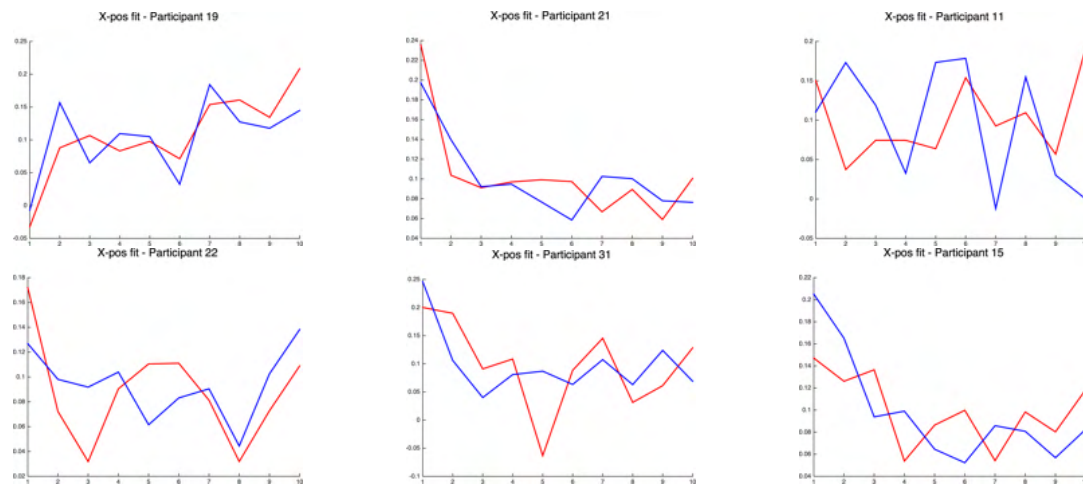


Figure 2.15: Examples of participant weights.

2.3.7 Follow-up: Univariate stimuli

A follow-up smaller study tested whether participants would exhibit behavior closer to the predicted optimal when the task was simpler, location only varied across one axis. The presentation duration was also manipulated to verify whether the overweighting in the dependent condition, would be affected in the same way as for the independent condition. Each dot was presented for 500ms before it disappeared and was followed by a 500ms inter-stimulus-interval (ISI). Due to the increased amount of time needed for each trial, only 250 trials were shown to participants to keep the expected experiment duration within an hour. Results (N=10) showed that while the primacy effect was eliminated for the independent condition, it was still present for correlated samples (see [Figure 2.16](#)).

2.3.8 Follow-up: Prediction task

In order to ensure that participants were sensitive to the difference in the statistical structure, a within-participant prediction task was conducted. The presentation of the stimuli was the same as for the previous follow-up, but participants were asked to predict the location of the next dot in the sequence.

Ten participants (9 across both experiment days, 7 females, aged between 21-27 years old) completed the task. Prediction weighting was consistent with the expectation that in the inde-

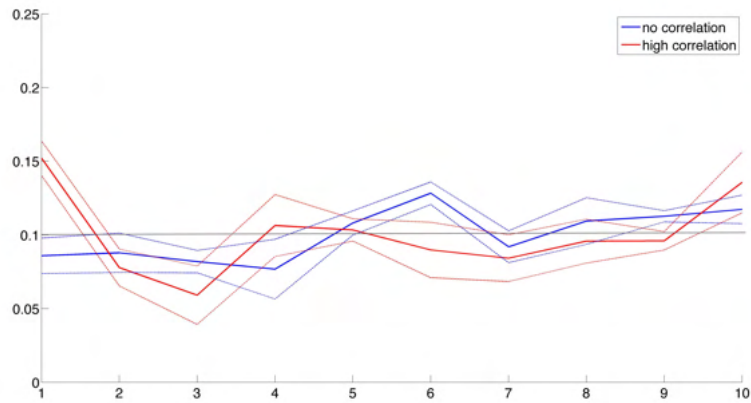


Figure 2.16: Average weights for estimation of the sequence center.

pendence condition samples should be weighted uniformly, and in the high dependence condition considerably more weight should be placed on the last sample/s. For all participants, the recency only model was the best fit in the dependent condition, as participants only relied on the last one or two stimuli in the sequence to make predictions (Figure 2.17). This reinforcing the fact that participants are sensitive to the change in dependence structure, and in anything, over-estimated the amount of dependence in the data. Therefore, a lack of sensitivity to the generative process is not the source of the heterogeneity observed in the estimation task.

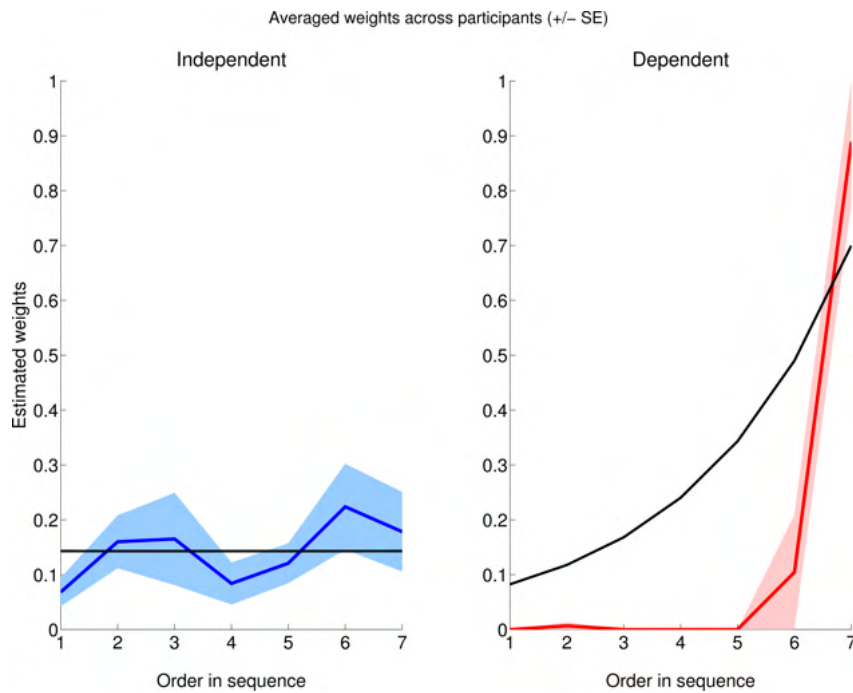


Figure 2.17: Average prediction weights.

2.3.9 Experiment 2

Methods

Participants

In total, 35 participants took part, 27 females, ages 19-25. Out of these, 32 returned to the lab to complete the experiment and were included in the analysis.

Design and materials

The task design closely matches that of Experiment 1. However, abstract shapes (covering a 1cm area) were used instead of the dots as stimuli. For each trial, 8 shapes were randomly selected (without replacement) from a pool of 26 shapes. [Figure 2.18](#) provides some examples of shapes.



Figure 2.18: Sample stimuli for the estimation task.

Given the increase in stimulus complexity, the presentation time and ISI were both decreased to 150 ms. Each session contained 300 trials.

The procedure and data analysis methods remained the same.

Results

The averaged weights for each condition ([Figure 2.19](#)) did not conform to the expected pattern of results, as both conditions show a profile consisted with uniform weighting. However, there was a large amount of heterogeneity across participants ([Figure 2.20](#)). The regression model fits the data reasonably well, with mean R^2 values of .72 ($SD = .19$). However, the estimated weights were fairly imprecise. In the independent condition, 24 out of 32 participants (75%) were best fit by the uniform weights model, 4 by the recency model and the remaining 4 were

best accounted for by the full weights model. In the dependent condition, 17 participants were described best by uniform weights, 3 by recency, 2 by primacy and 10 by the full weights model.

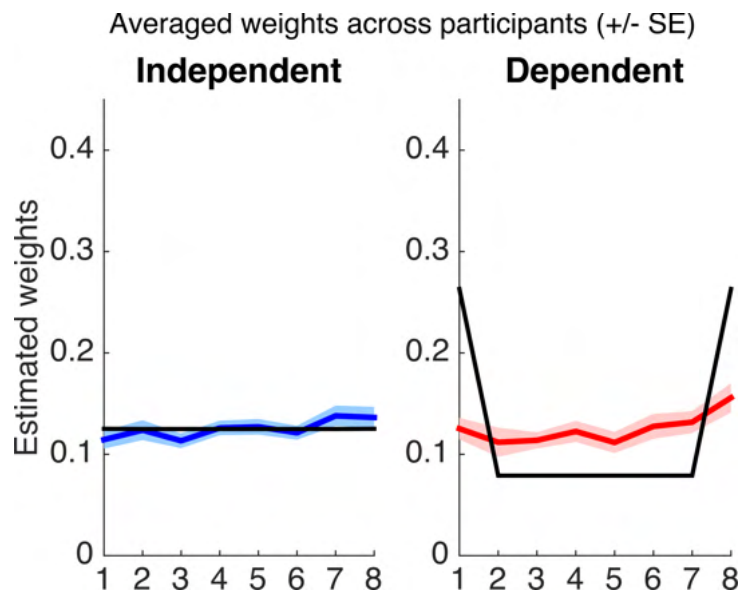


Figure 2.19: Estimated weights.

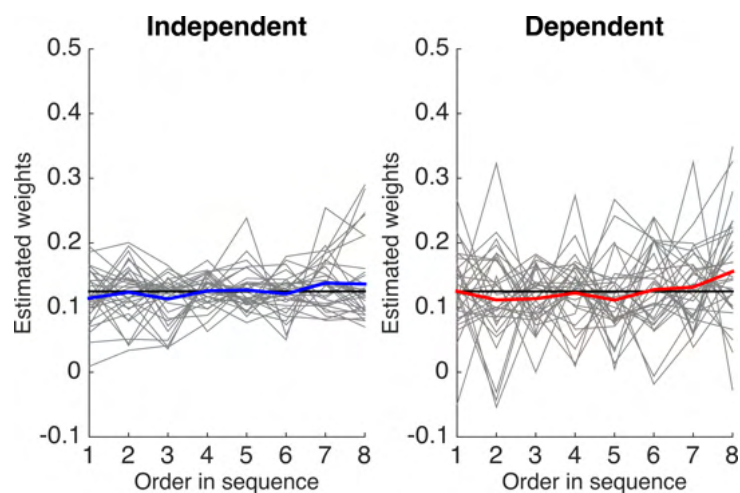


Figure 2.20: Estimated weights.

2.3.10 Follow-up: Memory task

Participants

Fifteen participants took part in the experiment, but only 11 completed all the four sessions

and are therefore included in the results. The age range of the group was 20-24 and 10 were female.

Design

Participants completed both the estimation task as well as a memory task. The estimation task was the same as presented above.

The memory task differed from the estimation task only at the response stage as within a trial they saw the exact same shapes identities, at the same locations as in the estimation task. Further, stimulus presentation times in the memory task coincided with those in the estimation task (Experiment 1). At the end of each trial, participants viewed 16 shapes presented centrally, half of which had been presented in the series, half of which were sampled randomly without replacement from a remaining pool of 18 shapes. They then judged whether they had seen them in the trial or not and responded by dragging the cursor to / clicking on a button which read yes/no. The trackpad was used for compatibility with the estimation task. Unfortunately, given the longer time needed to respond in the memory task, only 100 trials could be used per session to avoid fatigue. A full session of 250 trials was completed for the estimation task.

Procedure

There were four sessions in total, with the sessions of the same task scheduled on consecutive days. The order of the task and condition was fully counterbalanced.

Data analysis

Analysis for the estimation task was conducted as described above, but using only the first 100 trials shared with the memory task for task comparisons.

For the memory recall task, percent of shapes correctly remembered (hits) was calculated for every serial position and the percent of false alarms. The estimation weights and hits within a condition were then correlated for all participants.

Results

The U-shape serial order curve typical in recall tasks was observed for both conditions as can be seen in [Figure 2.21](#). In a 2x2 repeated ANOVA with serial position and condition as factors, the interaction between the two factors was non-significant, $F(7, 160) = .07$, $p = .99$.

, but the main effect of condition was significant, $F(1, 160) = 4.28, p = .04$, as performance in the independent condition was marginally better. For both conditions, primacy and recency were observed, pooled across conditions the first sample was significantly different from .5, $t(21) = 6.48, p < .001$, as well as the last one, $t(21) = 13.85, p < .001$.

The estimation performance matched the results of Experiment 1 (see Figure 2.22), and, as previously, was characterized by large inter-individual differences. There was also little convergence between estimation weights and memory weights as can be seen in Figure 2.23 which illustrate the relationship between them for two example participants. The correlation was low in both the independent, $.03, p = .82$, as well as dependent condition, $.08, p = .45$.

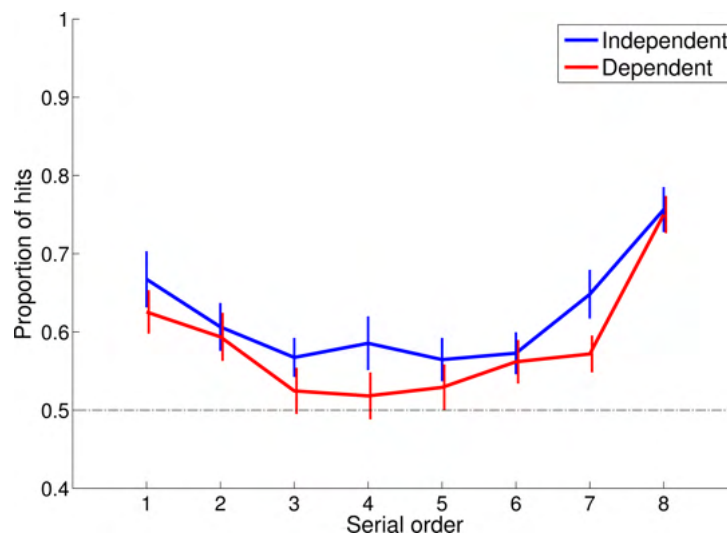


Figure 2.21: Mean recall performance by condition. Vertical segments mark SEs.

2.3.11 Conclusion

Optimal estimation from sequentially dependent, as opposed to independent samples, requires different strategies. Given the ubiquity and relevance of temporal correlations in the visual environment, if humans are to make decisions efficiently, they should exploit information about the correlational structure of sensory samples. We predicted that if adaptation to stimulus statistics is fast and flexible, people will be near-optimal in the estimation of summary statistics even in an unsupervised task. In this setting, the optimal incorporation of correlated samples should exhibit order effects - a symmetric overweighing of the first and last samples that is

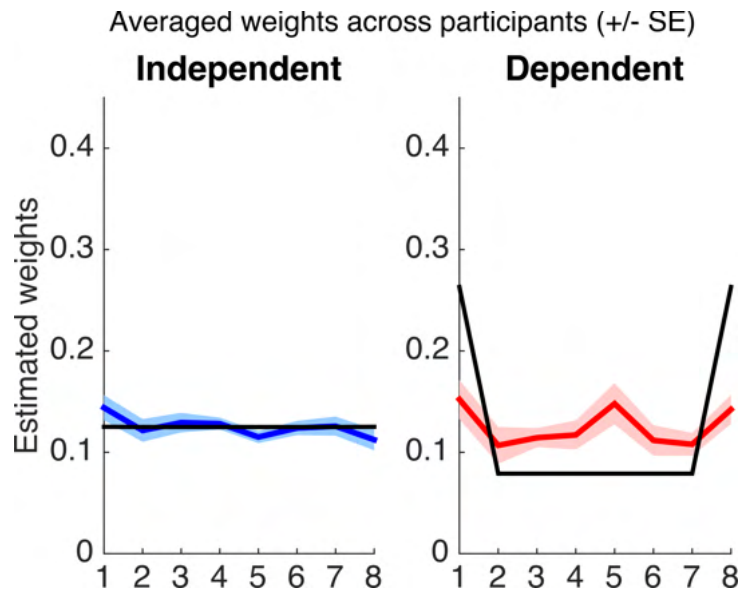
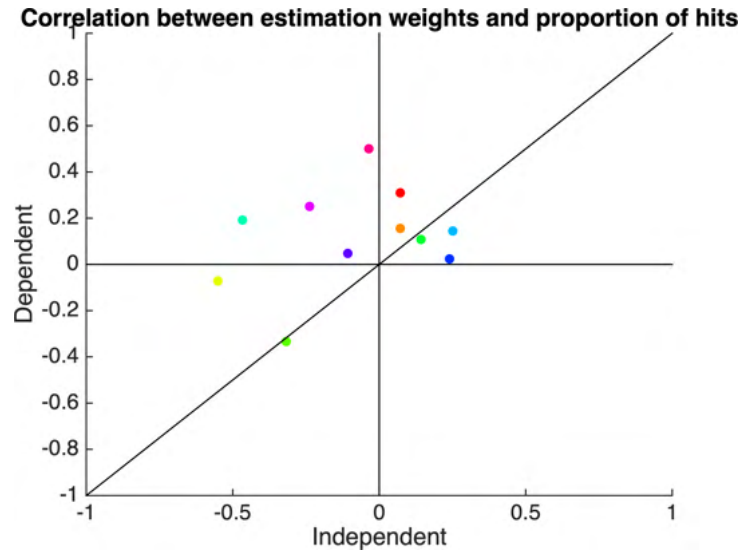


Figure 2.22: Average estimation weights.

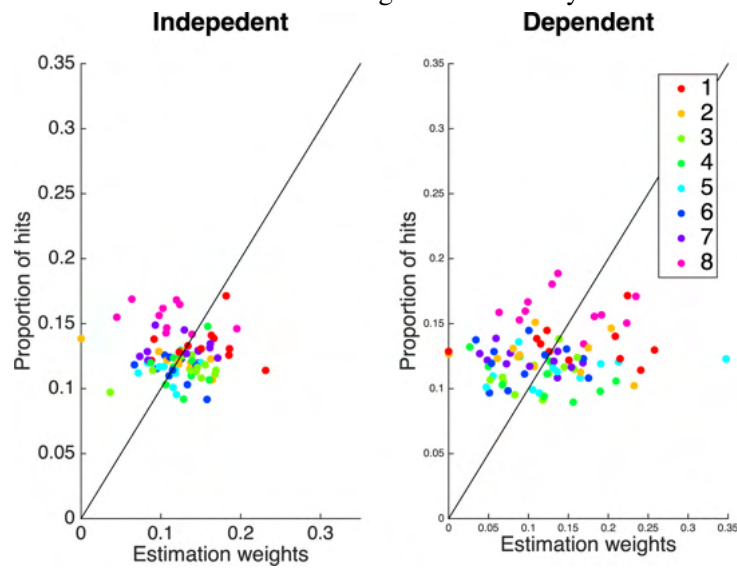
increases with the correlation in the data.

The multiple estimation experiments using stimuli characterized by one-dimensional or two-dimensional position and varying in stimulus identity have failed to demonstrate that humans exhibit a (near)-optimal, flexibly adaptive strategy in estimation. While aggregate results seemed promising, a very small number of participants showed both recency and primacy in the dependent condition, with the majority following a primacy only or a recency only weighting pattern. This pattern cannot be explained by noisy perception of the sample correlation, as participants were able to utilize the correlation structure when making predictions for future samples.

The results support for previous findings that early i.i.d. samples are more influential in mean estimation (Juni et al., 2010; Wallander & Boynton, 2013). However, based on the individual-level analysis, the group-level effects were also very likely driven by only a subset of participants. A large percentage of participants (around to 75% in our experiments) were following a uniform strategy even at short stimulus presentation times. Interestingly, while increasing the presentation time levelled out the weights in the independent condition as expected, this did not occur in the dependent condition. Further, while Wallander and Boynton (2013) found that predictability of the sequence decreased the early overweighting of evidence,



(a) Spearman correlation between estimation weights and memory recall within every participant.



(b) Relationship between the estimation weights and proportion of hits (hits at serial position out of total hits). Color corresponds to the serial position in the sequence and each dot represents a participant.

Figure 2.23: Relationship between serial effects in estimation weights and memory performance.

we found that early overweighting is more robust for structured sequences.

Lastly, we found that variability in estimation weights across serial positions was not related to memory recall. However, we unexpectedly found that recall for the independent condition was better than for the dependent condition. However, the effect size was very small.

It is possible that given the relatively small expected effect size, the task used was not sensitive enough to pick out crucial differences or sufficiently engaging for participants. Perhaps a semi-supervised task which in the loss accrued with sub-optimality is emphasized would lend itself better to the task.

2.4 General Discussion

Across the teaching and learning studies presented in this chapter, we found that participants are sensitive to sampling demands both when teaching and when learning. This was evidenced in the fact that teachers did not select examples according to random or strong sampling, and learners clearly distinguished autocorrelated from i.i.d. sequences. At the same time, at the individual level, learning was not optimal and neither was the teaching.

The first study replicated a task requiring the teaching of prototype and a rule-based concepts. Results were mixed. Findings were consistent with the predictions of Shafto et al. (2014) in the prototype teaching game. In our extension of rule-based teaching, participants did not attempt curriculum teaching as previously shown, and only few participants (about a third) were able to reach the optimal solution efficiently. Moreover, while the participants improved over time, we found no graded association between the number of examples used and the final teaching outcome (although participants in the unconstrained condition were more likely to reach the optimal solution due to choosing more examples).

Making the number of examples needed to teach the boundary explicit to the participants did not aid participants, contrary to our predictions. This is surprising particularly because the Constrained condition followed the Unconstrained condition, which should have provided participants with experience on the task and made it more likely for them to gain insight about

the optimal solution structure. While the current dataset was not collected for this purpose, we can speculate that the prospects of explicit training to improve teaching performance might not be particularly high.

It also stands to weaken alternative explanations for the original results in Khan et al. (2011) on 'over-teaching', that is choosing more examples when the teaching goal has already been achieved, such as insensitivity to teaching costs or a desire to show willingness to teach by the amount of effort put in. It seems plausible that teachers did not identify the minimally sufficient teaching set easily.

Simplifying the stimuli by using only one dimension, with easy to perceive differences and straightforward ranking did not lead to improvements in teaching performance. This is potentially indicative of the fact that an understanding of what makes a teaching set useful for a learner was the underlying difficulty during this task.

Lastly, how can we relate the findings of the two replications? Arguably, the task in Khan et al. (2011) was a simpler task in that the pedagogical sampling solution is deterministic and does not require making any pragmatic inferences about the learner. In other words, even a learner that does not know that they are being taught will succeed in the learning task if the optimal teaching set is provided.

It can be argued, however, that the results of Shafto et al. (2014) can be an effect of how the original information has been encoded (rather than of inference based on a learner's ability to exclude alternative hypotheses). The visual system is particularly adept at extracting summary statistics from stimulus ensembles (especially homogeneous ones), with the central tendency (mean) and range being pivotal, and there is evidence that identification of set members is heavily modulated by these summary statistics, as well as outliers. It would be, therefore, very useful to further corroborate the fact that participants are able to randomly and explicitly sample from the same distribution. This is a complicated experimental challenge since most requests for explicit information from participants could be regarded as communicative and perhaps pedagogical. The fact that repeated human 2AFC choices can be used to sample internal representations (analogous to Markov Chain Monte Carlo sampling) has been shown before

(A. Sanborn & Griffiths, 2007), but there is little work on explicit distributional expectations.

Other promising avenues are explicitly establishing within the experimental context the number of hypotheses and their relative location in the stimulus space and checking whether predictions based on these modulations closely follow experimental data.

Overall, these preliminary results warrant questions about the abilities to teach when more realistic demands are added such as producing the samples versus choosing samples from a predefined array, uncertainty about the boundaries, etc.

In the second study, learners did not aggregate information as a function of the serial auto-correlation to produce optimal estimates for the mean of the series. This was surprising given the clear need to adapt to correlated inputs generated by environment sampling or by other people. However, the possible improvement associated with the optimal strategy were potentially too low, so we suggest that current prediction might be more adequately tested by applying explicit costs or implicit ones (e.g., motor costs).

We know very little about learning adaptations to the kinds of structures and statistical patterns generated (intentionally but also unintentionally) by other humans when collaborating and teaching. Experimental work has uncovered expectations that others are rational (here, maximally informative) and violate strong sampling, but this research project is still in its infancy. It is still an open question how much of the range of human behavior these assumptions can explain.

2.5 Supplementary Information

2.5.1 Study 2: Additional figures

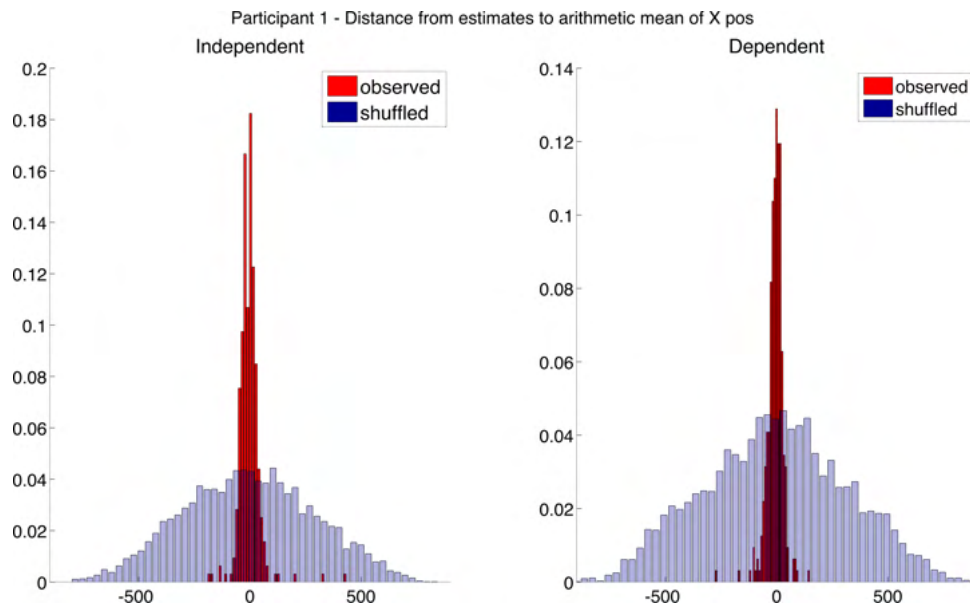


Figure 2.24: Estimation performance of Participant 1. The observed signed distances from arithmetic means of trials and the participant's estimates (red) against the distances between estimates and shuffled trial means.

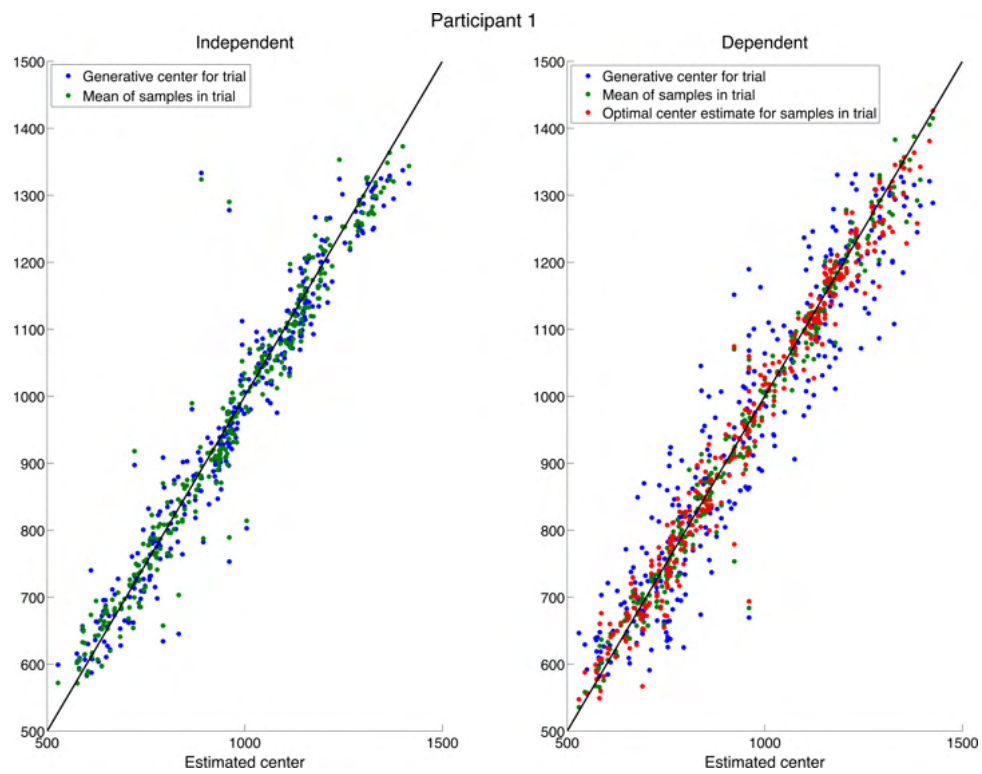


Figure 2.25: True center for trials against means computed using the optimal strategy and uniform weights.

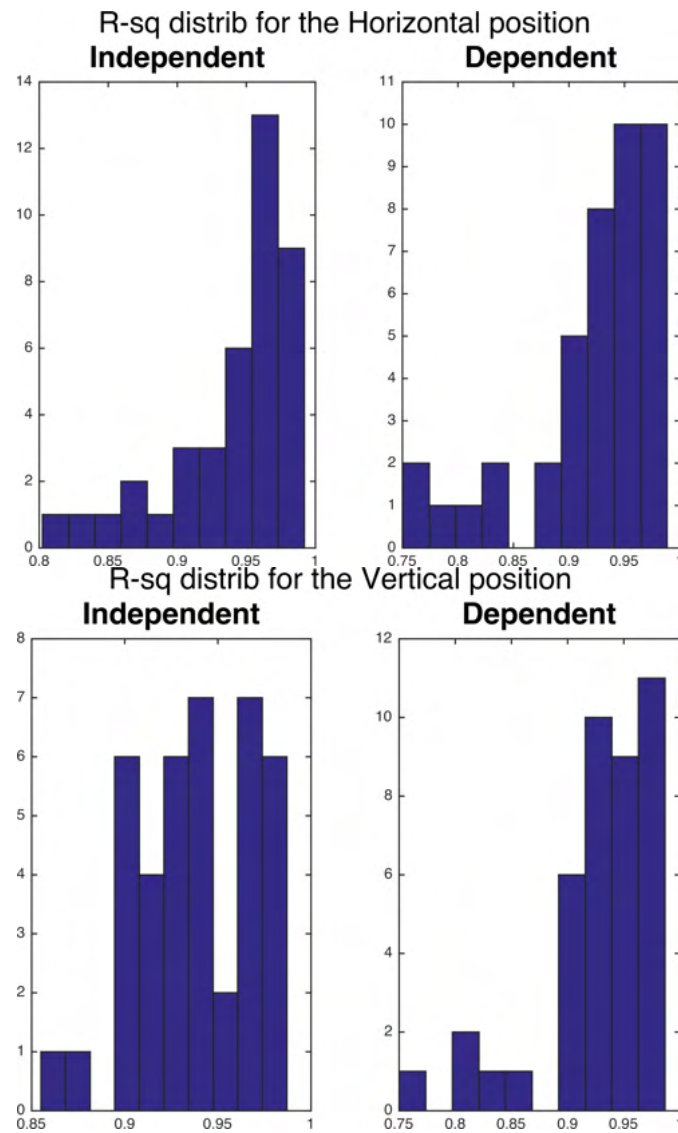


Figure 2.26: Variance explained (R^2)

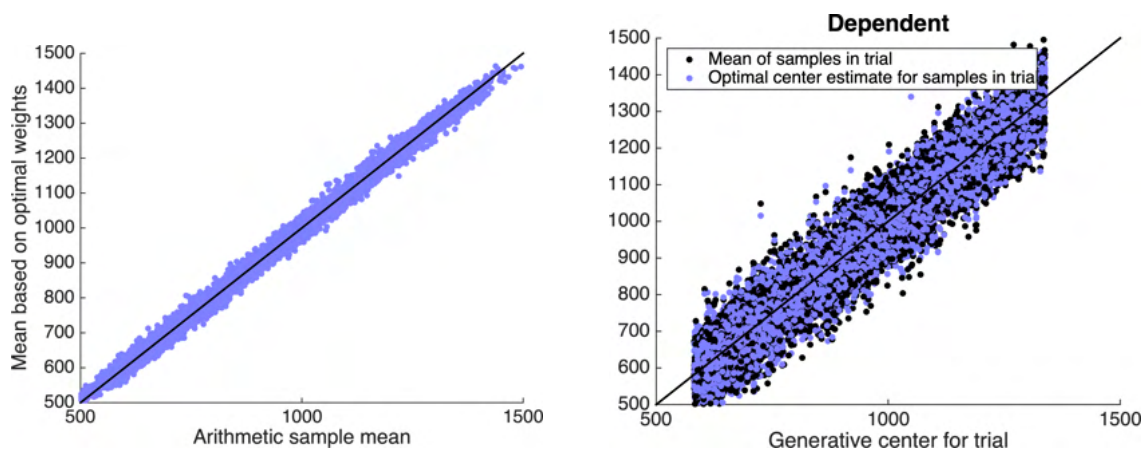


Figure 2.27: Predictions based on optimal and uniform weighting (Left); Ground truth relative to optimal and uniform weighting (Right).

AIC - Exp 1

| | INDEPENDENT CONDITION | | | | | | | | DEPENDENT CONDITION | | | | | | | |
|----|-----------------------|--------|--------|--------|----------------|--------|--------|--------|---------------------|--------|--------|----------------|--------|--------|--------|-------|
| ID | Uncons | Primac | Recenc | Symme | Asymmetric Pri | Linear | Quadr | Uncons | Primac | Recenc | Symme | Asymmetric Pri | Linear | Quadr | | |
| 1 | 4515.4 | 4501.2 | 4504.8 | 4505.2 | 4503 | 4502.8 | 4505.2 | 18432 | 4588.2 | 4576.1 | 4577.1 | 4577.9 | 4577.3 | 4575.5 | 4580.7 | 17772 |
| 2 | 4555 | 4545.5 | 4545.5 | 4545.8 | 4547.3 | 4543.8 | 4555.5 | 18472 | 4584.1 | 4572.4 | 4577.9 | 4578.6 | 4573.9 | 4576 | 4583.1 | 17531 |
| 3 | 4526.2 | 4512 | 4514.7 | 4515.4 | 4513.6 | 4512.8 | 4529.4 | 17266 | 4644.3 | 4634.6 | 4633.4 | 4635.7 | 4634.7 | 4633.1 | 4676.2 | 16128 |
| 4 | 4597.5 | 4586.7 | 4586.2 | 4586.9 | 4588.2 | 4584.7 | 4588.6 | 17765 | 4566.9 | 4552 | 4557.1 | 4558.1 | 4554 | 4555.4 | 4537.2 | 17580 |
| 5 | 4576.1 | 4563.2 | 4563.2 | 4563.6 | 4564.9 | 4561.6 | 4579.4 | 18277 | 4563 | 4551 | 4550.7 | 4551.7 | 4552.2 | 4549.8 | 4560.7 | 17816 |
| 6 | 4564.9 | 4553.6 | 4555.1 | 4555.7 | 4555.3 | 4553.2 | 4569.1 | 18686 | 4567.2 | 4559.3 | 4558.3 | 4559.4 | 4560.1 | 4557.7 | 4551.2 | 16507 |
| 7 | 4560.4 | 4547.1 | 4547.5 | 4547.8 | 4549 | 4545.5 | 4554.7 | 19115 | 4601.9 | 4590.8 | 4592.1 | 4592.4 | 4592.7 | 4590.1 | 4594.8 | 17618 |
| 8 | 4546.1 | 4536.3 | 4536 | 4536.4 | 4538 | 4534.4 | 4538.8 | 17784 | 4535.5 | 4521.5 | 4531.2 | 4532 | 4522.5 | 4529.3 | 4536 | 17226 |
| 9 | 4524.3 | 4509.9 | 4512.6 | 4512.8 | 4511.8 | 4510.6 | 4531.2 | 17598 | 4592.1 | 4580.3 | 4587.9 | 4588.9 | 4581.3 | 4586.2 | 4609 | 17388 |
| 10 | 4572.9 | 4562.5 | 4562 | 4562.8 | 4564 | 4560.7 | 4554.9 | 17321 | 4584.9 | 4572 | 4577.2 | 4578.1 | 4573.4 | 4575.4 | 4572 | 16661 |
| 11 | 4568.7 | 4557.6 | 4556.7 | 4557.4 | 4558.7 | 4555.6 | 4553.9 | 18085 | 4557.8 | 4544.7 | 4544.8 | 4545.9 | 4545.8 | 4543.9 | 4550.2 | 17250 |
| 12 | 4598.5 | 4587 | 4586.5 | 4587.1 | 4588.5 | 4585 | 4600.7 | 17379 | 4591.9 | 4577 | 4583.9 | 4584.3 | 4579 | 4582 | 4598.8 | 18003 |
| 13 | 4541 | 4528.2 | 4529 | 4529.4 | 4530.1 | 4527.1 | 4517.6 | 17625 | 4612.8 | 4602.6 | 4600.6 | 4603.6 | 4601.9 | 4601 | 4602 | 17426 |
| 14 | 4596 | 4584.2 | 4584.9 | 4584.9 | 4586.2 | 4583 | 4609.6 | 17312 | 4609.6 | 4601.2 | 4600 | 4601.2 | 4601.8 | 4599.6 | 4586.7 | 16701 |
| 15 | 4553.9 | 4547.4 | 4546.9 | 4547.4 | 4548.6 | 4545.6 | 4545.1 | 15573 | 4544.2 | 4530.4 | 4538 | 4538.8 | 4531.7 | 4536.1 | 4540 | 18665 |
| 16 | 4574.9 | 4562.9 | 4563.4 | 4563.3 | 4564.8 | 4561.4 | 4559.8 | 17982 | 4528.8 | 4515.2 | 4523.1 | 4524.1 | 4516 | 4521.3 | 4520.2 | 17188 |
| 17 | 4570.2 | 4557.9 | 4557.9 | 4558.4 | 4559.5 | 4556.4 | 4577.9 | 18247 | 4555.9 | 4542.2 | 4550.4 | 4551.1 | 4543.9 | 4548.4 | 4546.7 | 18103 |
| 18 | 4596.3 | 4584 | 4584 | 4584.4 | 4585.7 | 4582.4 | 4614.4 | 17584 | 4600.3 | 4588.3 | 4595.9 | 4596.7 | 4589.6 | 4594 | 4617.2 | 18006 |
| 19 | 4587.6 | 4573.4 | 4574 | 4574.3 | 4575.3 | 4572.2 | 4591.1 | 17428 | 4586.6 | 4574 | 4574.3 | 4575.3 | 4575.3 | 4573.3 | 4615.9 | 16984 |
| 20 | 4579 | 4571.9 | 4571.5 | 4571.8 | 4573.2 | 4570.1 | 4580.1 | 14978 | 4615.1 | 4604.6 | 4603.5 | 4605.9 | 4604 | 4603.4 | 4625.3 | 17414 |
| 21 | 4528.9 | 4518.7 | 4518.8 | 4519.2 | 4520.4 | 4517.2 | 4528.6 | 18837 | 4539.9 | 4534 | 4533 | 4534 | 4534.9 | 4532.2 | 4569 | 17284 |
| 22 | 3684.6 | 3675.5 | 3678.2 | 3680.1 | 3676.6 | 3678.1 | 3674.9 | 15032 | 3735.8 | 3723.3 | 3729.5 | 3731.5 | 3723.1 | 3728.6 | 3728.1 | 13958 |
| 23 | 3662 | 3655.2 | 3655.2 | 3655.9 | 3656.6 | 3654 | 3654.8 | 14970 | 3741 | 3728.5 | 3730.9 | 3732.4 | 3728.6 | 3729.7 | 3728 | 12517 |
| 24 | 3712.3 | 3700.3 | 3700.7 | 3700.6 | 3702.3 | 3698.7 | 3724.4 | 14931 | 3736.6 | 3727.1 | 3723.6 | 3727.9 | 3725.1 | 3725.3 | 3742.7 | 14392 |
| 25 | 3712.9 | 3701.8 | 3702.5 | 3702.4 | 3703.8 | 3700.5 | 3707.6 | 12365 | 3709.2 | 3696.6 | 3697.3 | 3697.6 | 3698.5 | 3695.4 | 3694.2 | 13549 |
| 26 | 3699.8 | 3692.9 | 3692.1 | 3692.8 | 3693.7 | 3691.2 | 3713.7 | 15191 | 3736.9 | 3727.5 | 3728.1 | 3728.6 | 3729.2 | 3726.3 | 3757.7 | 15033 |
| 27 | 3698.7 | 3687.6 | 3688.6 | 3688.9 | 3689.3 | 3686.7 | 3687.2 | 14665 | 3768.1 | 3756.6 | 3757.9 | 3758.4 | 3758.3 | 3756 | 3760.5 | 13464 |
| 28 | 3724.5 | 3714 | 3715.1 | 3715.4 | 3715.9 | 3713.1 | 3741.4 | 14199 | 3738.4 | 3726.4 | 3730.9 | 3732.6 | 3726.2 | 3729.8 | 3730.5 | 13597 |
| 29 | 3724 | 3711.6 | 3711.9 | 3712.4 | 3713.3 | 3710.1 | 3710.4 | 14446 | 3766.7 | 3751.8 | 3758.3 | 3758.9 | 3753.5 | 3756.4 | 3756.1 | 13589 |
| 30 | 3726.1 | 3714.5 | 3714.5 | 3714.5 | 3716.5 | 3712.5 | 3723.3 | 14816 | 3742.3 | 3728.6 | 3731.2 | 3731.5 | 3730.6 | 3729.2 | 3729.2 | 13672 |
| 31 | 3712.8 | 3698.9 | 3698.9 | 3699 | 3700.9 | 3696.9 | 3720.4 | 15035 | 3729.3 | 3715.4 | 3715.1 | 3715.6 | 3717.1 | 3713.4 | 3718.5 | 13853 |
| 32 | 3715.2 | 3702.9 | 3702.9 | 3702.9 | 3704.9 | 3700.9 | 3723 | 14335 | 3767.3 | 3753.6 | 3753.5 | 3753.8 | 3755.4 | 3751.8 | 3771.5 | 12660 |
| 33 | 3710.1 | 3698.1 | 3699.8 | 3700.4 | 3699.9 | 3698.2 | 3716.3 | 15222 | 3704.2 | 3691.9 | 3692.5 | 3693.3 | 3693.3 | 3691 | 3691.2 | 14050 |
| 34 | 3739.4 | 3729.5 | 3729.6 | 3730.6 | 3730.4 | 3728.4 | 3728.1 | 14239 | 3690.3 | 3676.5 | 3681.3 | 3682.4 | 3677 | 3680 | 3688.8 | 14056 |
| 35 | 3706 | 3696.7 | 3691.6 | 3697.4 | 3693.6 | 3694.7 | 3691.5 | 14825 | 3724.3 | 3712.9 | 3715 | 3715.6 | 3714.7 | 3713.5 | 3707.8 | 13357 |
| 36 | 3705.3 | 3692.7 | 3693.3 | 3693.8 | 3694.3 | 3691.6 | 3711.9 | 14294 | 3729.8 | 3717 | 3716.5 | 3717.2 | 3718.1 | 3715.1 | 3727.8 | 13216 |
| 37 | 3758.3 | 3744 | 3744.6 | 3744.5 | 3745.5 | 3742.9 | 3754.1 | 14344 | 3739.3 | 3725.1 | 3728.3 | 3729.8 | 3725.6 | 3727.3 | 3737.7 | 13363 |
| 38 | 3658.7 | 3647.8 | 3647.1 | 3648.2 | 3648.8 | 3646.3 | 3651.1 | 13366 | 3712.9 | 3699.5 | 3704.8 | 3705.1 | 3701.3 | 3702.8 | 3704.2 | 13014 |
| 39 | 3710.8 | 3698 | 3697.4 | 3698 | 3699.4 | 3696.1 | 3700.8 | 13822 | 3706.3 | 3693.3 | 3692.4 | 3693.7 | 3694.2 | 3691.6 | 3692.9 | 12597 |

Figure 2.28: AIC values for the models fit in Experiment 1. Each row corresponds to a participant.

Chapter 3: Actively Learning How to Teach

3.1 Introduction

Perhaps the most enduring debate in the education literature, as well as in kindergartens and classrooms, concerns the virtues of exploratory play in contrast to the canonical, largely passive mode of teacher-led instruction (Bruner, 1961; Mayer, 2004; Montessori & Gutek, 2004). The discussion has been naturally phrased in terms of the relative benefits and disadvantages that usually young learners incur when learning from self-guided discovery compared to direct instruction. However, the complementary link between efficient self-guided learning and good teaching has not been explored thus far.

As outlined in the Introduction, teaching is difficult because it requires not just having a good grasp of the concept to be taught but also knowing the utility of the information that can be transmitted from the perspective of a potential learner. In the specific case of teaching by offering examples, there are general principles that ought to guarantee, in theory, the informativeness of a given set of examples for any possible learner: removing redundancies from the information provided, making sure that examples are not ambiguous with respect to the hypothesis to be taught. However, in addition to this, the usefulness of a piece of information depends on a myriad of interacting factors that govern how a learner would make use of the information provided: the context in which it is acquired (i.e. social cooperative context or not), when it is received, the learner's prior knowledge, and idiosyncrasies in how the learner makes inferences based on new data. Therefore, what example is most useful needs to be redefined in terms of the extended learning model.

The easiest way to build naïve models for learning on specific tasks (e.g. learning how to categorize items), or models of individual learners for the purpose of teaching, is to simulate or actually engage in the experience of being a learner.

Intuitively, taking the perspective of the learner prior to teaching should be a useful experience. It could allow the teacher to better understand, implicitly and/or explicitly, how a learner makes inferences to solve the task at hand based on the data provided, i.e. which data points were informative and which were not. For instance, it would allow the teacher to become aware of what kinds of hypotheses are warranted by specific data sets and perhaps partially eliminate ‘the curse of knowledge’. Instances of this effect are well-documented and widely spread, from failures to use sufficiently disambiguating language (Keysar & Henly, 2002) when addressing others who are less informed to failures to see how the same data visualization may be interpreted differently by others (Xiong et al., 2020).

Related to this, following learning, a teacher might more likely take into account the multiple hypotheses considered throughout the learning task, as opposed to just generating examples from the hypothesis they wish to teach, which although consistent with the data, would not necessarily disambiguate the hypothesis-to-be-taught from others.

Having the experience of being an active learner prior to teaching should generate particularly robust insights about how to select good examples for teaching, beyond what experience being a passive learner could offer. The commonalities between teaching and active learning are easy to identify when comparing their formal descriptions. Recent rational-agent models have conceptualized teaching as a recursive process in which the teacher and the learner reason about each other. Specifically, the teacher selects training samples for the learner such that, given the learner’s prior knowledge and inference making mechanisms, these samples would lead the learner to the desired conclusion efficiently, i.e. by requiring the smallest number of samples (Shafto et al., 2014). Conversely, the learner interprets the observed samples assuming they were generated by this pedagogical process (as opposed to randomly). Similarly, an ideal active learner will also sample the environment strategically. However, they will sample by directing their information gathering (e.g. by moving their eyes to explore a visual scene or

choosing interventions on the environment) in order to maximize their expected information gain (Yang, Wolpert, et al., 2018). There are two ways in which active learning can be advantageous. First, observations collected in a strategic way will be more informative for any learner (not just the one sampling information); for instance, by avoiding irrelevant or redundant evidence. Second, and more importantly, there is an added advantage specific to the active learner stemming from the fact that they sample information in light of their prior knowledge and the hypotheses that they wish to test. This effect was demonstrated in experiments where the data selected by an active learner were also presented to a yoked "passive" learner, and, despite the observations being matched, active learners performed better than their yoked passive counterparts (Markant & Gureckis, 2014).

Thus, both being a good teacher and being a good active learner rest on the same general ability to evaluate the potential value of a new piece of evidence relative to a current state of knowledge and a task. Nonetheless, there is an important difference. The active learner does not have access to the target hypothesis and thus can only select data that minimize uncertainty. However, Yang et al. (2019) more recently proposed a re-conceptualization of active learning as self-teaching by envisioning a learner who simulates an uninformed teacher whose task is limited to providing queries. In this framework, the self-teacher does not optimize for expected information gain, although this will often be the collateral result.

Another important caveat is that the teaching problem as stated is highly reliant on Theory of Mind (ToM) abilities, dynamically considering the belief states of another person. In line with this, Bass, Shafto, et al. (2017) have linked Theory of Mind (ToM) development to children's pedagogical sampling ability. While taking the role of the learner would encourage perspective taking, it might still only allow learners-as-teachers to imagine learners who have very similar learning trajectories to themselves.

Alternatively, if teaching strategies are highly automatic and follow stereotyped solutions derived from rational (pragmatic) accounts of teaching, it is possible that the mode of learning does not influence teaching performance.

While it is more immediate to consider if active learning could improve teaching, this rea-

soning can also be applied in the opposite direction to explain the protégé effect - improved learning through teaching. For instance, Bargh and Schul (1980) have shown that merely being told that one needs to learn to teach (as opposed to learning to sit a test) leads to better learning outcomes. The protégé effect motivates the use of peer-to-peer instruction and small unit work groups in classrooms, as children tend to enhance their own learning through mentoring others. A leading cause of this effect is undoubtedly increased engagement in the learning process and more emotional involvement. While difficult to test experimentally, it is likely that benefits of teaching for learning supersede the improved engagement. Particularly, it is likely that peer teachers are forced to be reflexive about their reasoning by deciding which information to present and they are likely to become more aware of any knowledge gaps due to having to answer their mentees' questions. In a series of classroom studies, Leelawong and Biswas (2008) tested differences in learning between school children who, over an extended period of time: were taught, learned by teaching a virtual Teachable Agent simulating a passive tutee, and a virtual agent that self-regulated (simulating an active peer). Students who taught over-performed those who were taught, and the students paired with Teachable Agents prompting metacognitive assessments were best, but benefits of teaching dwindled for a novel transfer task (students did not develop long lasting learning behaviors). Lastly, Shafto et al. (2014) found that learners were more likely to conform to the predictions of the rational teaching model if they first had teaching experience.

There are further interesting and ecologically relevant questions about the interplay between active learning and teaching, particularly as current best practices in education emphasize the use of student-centered approaches (EHEA, 2020; Klemencic et al., 2020) for learning and assessment. The student-centered approach translates into a form of semi-supervised learning blending active learning with tailored guidance from teachers. While the ability of children to engage in exploration is undoubtedly a primary contributor to the benefits of such modes of education, it would be interesting to quantify the role of teaching in these settings. Specifically, the amount of information that a teacher can glean about a learner should differ as a function of whether their student is learning actively or passively. Specifically, active learning settings

should allow teachers to infer the uncertainty of the learner more reliably which can in turn help them tailor the information they provide. Therefore, it is possible that a considerable boost to the performance of learners in such environments is actually driven by the better instruction stemming from teachers having a better model of their active students.

Further, while there is ample evidence of the benefits of active exploration relative to passive learning, there are tasks for which active learning does not supersede supervised learning in terms of speed of acquisition. For instance, Markant and Gureckis (2014) found poor active learning performance for a categorization task in which the boundary depended on two perceptual features, on par with unsupervised learning. However, using the same categorization task with supervised (feedback) learning, Ashby et al. (1999) found that all their participants were able to learn the integration based boundary. These results highlight the existence of an significant potential bottleneck for active learning. Markant and Gureckis (2014) explain the performance of active learners through a sequential hypothesis dependent testing model according to which participants sample examples close to their current subjective boundaries, which, in this case, are predominantly unidimensional, leading to the observed pattern of poor performance. This highlights the fact that active learning alone is not sufficient in situations in which the hyperpriors dictating the hypotheses entertained by the learner are not aligned to the true structure of the stimuli. One of the important roles of teaching in this situation could be to guide the hypothesis generation process of learners.

To summarize, there are clear theoretical parallels between active learning and teaching, and very ecologically relevant questions about the dynamics between active learners and teachers which are currently unexplored through controlled empirical studies. In the following section, we propose two experiments that begin to explore this direction and we propose further extensions.

3.2 Study motivation

Given the computational similarity of teaching and active learning, is it possible that they are also integrated through linked processes in human behavior? In other words, would it be possible to hone teaching skills through active learning? If both tasks rely on a core ability to sample environmental data efficiently, the transfer could occur automatically during learning, without the knowledge or expectation that the acquired information will need to be used for teaching in the future. Furthermore, active learning should improve teaching performance beyond passive learning (even when the same information content is acquired) if the active selection of data is the crucial driver of the learning effect, rather than the benefit of familiarity with the teaching task or taking the perspective of a learner.

Both Experiment 1 and 2 directly tested the hypothesis that active learning experience improves teaching. In Experiment 1, the teaching performance of participants who first had active learning experience on the same task was compared to the teaching performance of participants who were yoked passive learners or had no learning experience. Similarly, in Experiment 2, teachers who were active learners were compared with teachers who were yoked passive learners.

Across both experiments, we tested whether the observed differences could be confounded by differences in accuracy on the learning tasks. Further, the extent to which active learning performance was (directly) predictive of teaching performance was investigated.

The tasks used in the two experiments were conceptually similar as they involved the teaching of a categorization boundary. While certainly, there are far more ecologically relevant tasks for the problems of everyday teaching, categorization was used because it has a well established tradition of study, specifically of allowing for precise quantification of learning as a function of the way in which stimuli are presented and the (potential) supervision is offered.

In Experiment 1, participants were tasked with teaching an arbitrary deterministic one-dimensional boundary. All possible stimuli (for active learning and teaching) were already ordered and continually visible to the participant. This task has a straightforward optimal so-

lution for teaching and active learning. Despite the conceptual simplicity of the problem and the scaffolding offered by its visual presentation, previous work has provided mixed evidence about participants' ability to reach an optimal solution in tasks with a similar structure. While Shafto et al. (2014) found evidence consistent with optimal behavior, Khan et al. (2011) found that participants use a variety of strategies and often do not discover the optimal teaching strategy in a more ecological setting. While a task similar to the ones previously used in teaching experiments was chosen to be able to relate our results directly to previous findings, it has one considerable limitation for testing the current hypothesis. Namely, this since the task can be thought of as an insight problem and the optimal solution is easily verbalizable, it is highly likely that any transfers between active learning and teaching were explicit.

Experiment 2 aimed to replicate the findings of Experiment 1 using a task that was richer and more fit for purpose. Participants were still required to teach a deterministic categorization boundary, but this time it was defined in a perceptual (bivariate) stimulus space. The critical difference is that the task required teachers to first learn across time the perceptual categories to be taught, either actively or passively. This is a meaningful departure in several regards that more closely resemble real-life teaching: the teaching strategy is very difficult to verbalize and transfer explicitly, the teacher may hold uncertainty in their beliefs about the hypothesis-to-be-taught and potential learners' representation is noisy. To our knowledge, this is the first teaching task used in which the learning of the material to be taught was done over an extended period of time (largely, previous experiments use simultaneous presentation of entire stimulus space). Further, participants designed their own stimuli to offer as examples, rather than selecting them from a preset array.

Beyond providing a replication to Experiment 1, the second task was intended to be a good testbed for future investigations of the adaptability of teaching behavior as it would allow for tracking changing representations across time, and well as manipulation of the structure of the categories to be learned which could result in different priors over potential hypotheses considered and different optimal teaching sequences. The specific learning task has been taken from Markant and Gureckis (2014) who exploited a well studied distinction in the categorization lit-

erature between rule-based and information integration based categories¹, which are thought to be situated on different sides of the explicit-implicit learning divide, to reveal the use of sequential hypothesis generation as the basis of query generation during active learning. The category structure distinction is also interesting from the perspective of teaching. At a basic level, we can inquire whether the category structure modulates teaching behaviour as predicted by the rational pedagogical model or whether a general heuristic is used for all teaching problems in this stimulus space. Further, since Markant and Gureckis (2014) revealed that participants had an early (and protracted) predisposition for generating queries that tested rule-based hypotheses during active learning of both categories. Given this, we can ask whether teachers attempting to teach a rule-based category provided examples in light of considering only rule-based boundaries as potential hypotheses or whether they still took into account the wider range of possible category structures (that were not encountered in their direct learning experience). Of course, in case they do not consider the full range of possible hypotheses it could be either because they failed to generate the full space of hypotheses or it could be that their experience makes them believe that they are unlikely to be entertained by another learner.

Despite the increased complexity, the second task is also amenable to predictions based on the rational pedagogy model. This model makes qualitatively different predictions for the best teaching sets based on the category structure, but also based on the teacher's uncertainty about the true hypothesis. Specifically, during the categorization task following learning, participants exhibited various levels of stochasticity in their responses in relation to the boundary. Teaching of a boundary needs to balance two conflicting goals: offering examples as close enough to the boundary to be able to maximally constrain the set of possible boundaries for the learner to infer, and not providing examples that are so close that the teacher, due to imperfect knowledge, accidentally mislabels the examples and misleads the learner. To the extent that the teacher's estimate of the boundary is more uncertain, if they are aware of this, they should be more cautious and provide teaching sets further away from the subjective boundary.

¹The boundaries were deterministic for both the rule-based and information integration structures. Here, rule-based denotes unidimensional rules (generally, rules that can be described in Boolean algebra) and integration-based denotes two-dimensional rules (cannot be described in Boolean algebra, categories are learned implicitly by participants and they cannot verbalize their strategies.)

Lastly, while in the current design teachers were choosing examples for an imagined learner, the task presented here could be used, with minor modifications, for closed loop experiments in which learners and teachers interact. In such a setup, experiments could test the extent to which teachers who are allowed to observe a specific learner are able to tailor the teaching sets to the idiosyncrasies of the learner and the extent to which the learners benefit from tailored real-time feedback versus a generic presentation of idealized teaching sequences.

3.3 Experiment 1

3.3.1 Introduction

In order to test the hypothesis that active learning improves teaching performance, we designed a simple task in which participants were required to both learn a one-dimensional categorization boundary over sets of objects, and teach it, in counterbalanced order. Crucially, different items and boundaries were used for the learning and teaching phases.

There were two independent groups of participants in our design, those who learned actively first and then taught, and those who first taught and then performed active learning. In addition, to probe whether the effect learning on teaching performance was specific to active learning, a yoked control group performed the same teaching task after learning passively from watching the active learners' labeled queries. We contrasted the teaching performance of these three groups in terms of the expected performance of a learner based on their teaching set.

3.3.2 Methods

Participants

Eighty-eight participants² (54 female, $M_{\text{age}} = 24$ years, range = 18 - 42 years old) were recruited from the local population through the university online research participation system and the

²A sample of 30 participants was planned per condition. A participant and their assigned dyad pair were excluded from the experiment due to poor English language comprehension.

local student union. Fluency in English was a precondition for participation and Hungarian speaking participants were additionally provided with Hungarian translations of the instructions. Participants were remunerated either in cash or supermarket vouchers to the value of 1,500 HUF (approximately €5 at the time). Ethical approval for the experiment was obtained from the United Ethical Review Committee for Research in Psychology (EPKEB) in Hungary.

Tasks

All tasks (active learning, passive learning, and teaching) consisted of three trials. In each trial, participants were shown eight images in a horizontal array such as the one displayed in [Figure 3.1](#). Participants were told that the images were sorted left-to-right according to a given "key" feature. For instance, animals were sorted according to their speed relative to body size or the average amount of time they sleep, or foods were sorted by their carbon footprint or their vitamin content. Images belonged to one of two categories (which were clearly marked at the extremes of the image array) according to whether their key feature was below or above a "boundary" (threshold value) which lied between two adjacent images (i.e. at one of seven possible locations). Unknown to the participants, the true boundary locations which dictated the category membership of the images were uniformly sampled in each trial from all the possible locations. Participants were provided with a description of a seemingly objective classification boundary (e.g. that slow and fast animals were separated by the speed of the average human scaled by size). These descriptions were intentionally chosen such that the participants were unlikely to have any strong priors about the location of the boundary (see Supplementary Information subsection [3.6.1](#) on Piloting Priors about the boundary location). Knowing the participants' prior was essential because it determined the optimal query choice in active learning.

The categories used for the learning and teaching tasks were randomly selected for each participant. Images and category cover stories were only presented once throughout the entire experiment.

In **active learning trials**, participants first saw the image array alongside the description



Figure 3.1: Example image array from the teaching task. In this trial, food items were sorted from left to right in ascending order of their vitamin B content. The black vertical bar represents the daily recommended dose of vitamin B, which is the boundary the participant had to teach. In this case, the participant clicked on the two images closest to the boundary, which were automatically labelled.

of the categories and the boundary, following which they were told that their task was to find the boundary by querying two images. An image could be queried by clicking on it, which immediately revealed its category membership through the color of the frame drawn around it. After the second query, participants were asked to pinpoint where they thought the boundary was located, again by clicking on one of the seven possible boundaries. Participants received feedback on whether they were correct, unlucky (they selected a boundary compatible with the labelled images that was not the true boundary) or incorrect (selected an incompatible boundary).

The **passive learning trials** had the same structure, except that the labels of two images were sequentially revealed to the participants before they had to make their decision about the location of the boundary. Crucially, for each passive learning participant, the images labelled corresponded to the queries of a previous active learning participant.

In **teaching trials** (see [Figure 3.1](#)), participants were shown the boundary separating the two categories and were asked to teach it to another participant who they were told would take part in the experiment at a later time. It was made explicit that the other participant would be presented with the same set of sorted images. The participant only needed to click on an image to mark it as an example, and it was automatically labelled. Mirroring the learning tasks, participants were only allowed to provide two examples, which is the number of examples sufficient to fully specify the correct boundary. Intuitively, selecting two adjacent images with different labels is sufficient to identify the boundary in this task.

Materials

All the images were selected from the MultiPic databank of standardized color drawings of concrete concepts (Duñabeitia et al., 2018).

Procedure

Participants were randomly assigned to one of three groups: active learning followed by teaching (N = 29), passive learning followed by teaching (N = 29), and teaching followed by active learning (N = 30). The experiment was presented on a 27inch screen in a quiet room and lasted for an average of 20 minutes (unspeeded). Following the experiment, participants completed an open-ended questionnaire about the strategies that they used to solve the tasks.

Quantifying performance

Teaching performance was measured by the information gain, IG_{teach} , which is the amount of entropy by which the teacher reduced the imagined learner's prior entropy $\mathbb{H}(b)$ by labelling two images:

$$IG_{\text{teach}} = \mathbb{H}(b) - \mathbb{H}(b|s_1, s_2, l_1, l_2) \quad (3.1)$$

where s , l , and b respectively denote image stimuli, category labels, and potential boundary locations. \mathbb{H} is the Shannon entropy over the possible hypotheses, the prior entropy is $\mathbb{H}(b) = -\sum_{b \in \mathcal{B}} P(b) \log_2 \frac{1}{P(b)}$, where $P(b)$, the learner's prior over the boundary locations, is assumed to be uniform. The optimal teaching strategy is to label the examples immediately preceding and following the boundary as this will eliminate any uncertainty about the location of the boundary, thus reducing all of the original entropy. On the other hand, selecting an example set that will leave the learner uncertain about the true hypothesis because many potential boundaries compatible with the example set will translate into a lower information gain.

Using the observed information gain to evaluate active learning performance would intro-

duce arbitrariness since it cannot distinguish a learner's well-planned query from a lucky one. An ideal learner should choose a query in light of their uncertainty about the labels that will be observed. First, learners should compute the expected information gain of the queries by weighing the posterior entropy by the probability of observing the given labels for the query made and then choose the query that maximizes the expected gain. Therefore, $\text{EIG}_{\text{learn}}$, the sum of the expected information gain of the first and second queries, was used instead of observed information gain. The expected information gain of the first query is:

$$\text{EIG}_{\text{learn}}(s_1) = \mathbb{H}(b) - \sum_{l_1 \in \mathcal{L}} \mathbb{H}(b|s_1, l_1) \cdot \sum_{b \in \mathcal{B}} P(l_1|s_1, b) P(b) \quad (3.2)$$

After observing the first label, the prior over the boundary locations is updated, and the expected information gain is computed again relative to the entropy remaining after the first labelled sample:

$$\text{EIG}_{\text{learn}}(s_2|s_1) = \mathbb{H}(b|s_1, l_1) - \sum_{l_2 \in \mathcal{L}} \mathbb{H}(b|s_2, l_2, s_1, l_1) \cdot \sum_{b \in \mathcal{B}} P(l_2|s_2, s_1, l_1, b) P(b) \quad (3.3)$$

Unless otherwise specified, statistical analyses of participants' responses were performed based on the average measures of IG_{teach} and $\text{EIG}_{\text{learn}}$ in the three trials of each task.

Decisions about the boundary location

In learning trials, after observing two labelled stimuli, participants marked the location of the categorization boundary. Their choice could be assessed based on whether or not the selected boundary was compatible with the labelled images they had seen. However, simply using the proportion of compatible answers (across the three trials) to assess their performance ignores the fact that trials differed in the number of remaining compatible boundaries. To control for this and characterize performance appropriately, we fitted a model that captured the intuition that participants behaved optimally and selected (randomly) from among the remaining compatible boundary locations in some r fraction of trials, while in the rest of the trials they “lapsed”

and selected a boundary randomly among all locations:

$$P(\text{choice} = b_i | s_1, s_2, l_1, l_2) = r \cdot \mathbb{1}\{b_i \in \mathcal{B}_{\text{compatible}}^{(i)}\} \cdot \frac{1}{|\mathcal{B}_{\text{compatible}}^{(i)}|} + (1 - r) \cdot \frac{1}{|\mathcal{B}|} \quad (3.4)$$

Thus, $r = 1$ indicates optimal behavior, while $r = 0$ indicates chance performance. We estimated r for each participant by maximum likelihood (under the assumption that trials were *i.i.d.*). Plots presenting the relationship between the different measures of decision performance are presented in Supplementary Information 3.6.2.

Data analysis

Predictions were directly tested using planned independent t-tests to compare the teaching information gain in the teaching first and learning first conditions. Paired comparisons were used for the two groups who experienced being learners first, the active learners and passive learners. Correlations were calculated between learning and teaching performance.

Post-hoc analyses were conducted to ensure that variables extraneous to the predictions (e.g. stimuli) did not have a meaningful impact on performance or modulate the reported effects. The design of the experiment lends itself naturally to mixed model analysis, since it allows fitting trial level data (without aggregation) and can describe variation arising from the experimental design. Starting from a baseline fixed effects only model with the experimental condition as a predictor of teaching performance, we sequentially fitted and compared models using two additional fixed factors, learning performance and trial number (and their interactions with the condition), as well as random intercepts for participant and trial identity (i.e. dimension used for classification of the objects). Fixed effects were tested using log-likelihood ratio tests for nested models with the same random effects structure. Non-nested models were compared using the Bayesian Information Criterion (BIC) and Akaike Information Criterion (AIC). Similarly, random effects (fitted via maximum likelihood) were tested using log-likelihood ratio tests while keeping the fixed effects model identical. Given that the mixed-effects analysis confirmed the results of the planned comparisons on the aggregated trial data, we will focus on

these comparisons in the Results section for brevity and clarity.

3.3.3 Results

Descriptives

Despite the surface level simplicity of the teaching task, a large proportion of participants ($\approx 60\%$) did not perform it optimally (i.e. did not choose the two images on either side of the boundary as the teaching samples). However, prior active learning made it easier to gain insight into the optimal solution for teaching. More than half of active learners, 17 out of 29 participants, performed at ceiling level by consistently (across the 3 trials) providing example sets compatible with only one categorization boundary. In contrast, only 11 of 29 participants in the yoked passive learning group, and 7 out of 30 of the participants who did not complete a learning task before teaching managed to select the optimal example sets. The optimality of active learning performance is discussed in Supplementary Information 3.6.2.

Teaching performance across conditions

Confirming the main prediction, participants who were active learners before being teachers outperformed those who started directly with teaching, on average providing .63 bits, 95% CI [.22, 1.05], of additional information to their (fictitious) learners (see Figure 3.2). The group difference was highly significant in an independent t-test, $t(57) = 3.04$, $p = .01$, Bayes Factor (BF)³ = 10.81 in favor of the alternative hypothesis.

Learning passively before teaching conferred a smaller, but still significant, benefit relative to foregoing learning. Passive learning increased teaching information gain by an average of .45 bits, 95% CI [.05, .85], $t(57) = 2.26$, $p = .03$, $BF = 2.16$ in favor of the alternative.

While we found strong evidence in support of the differences between the groups completing the learning and teaching tasks in different orders, a possible concern was that these differences were not induced by the experimental manipulation *per se*. Specifically, if there are prior

³Bayes Factors were calculated for a null model that assumes a zero standardized difference between groups, and a Cauchy alternative with a prior scaled to an effect size of .7, following Rouder et al. (2009).

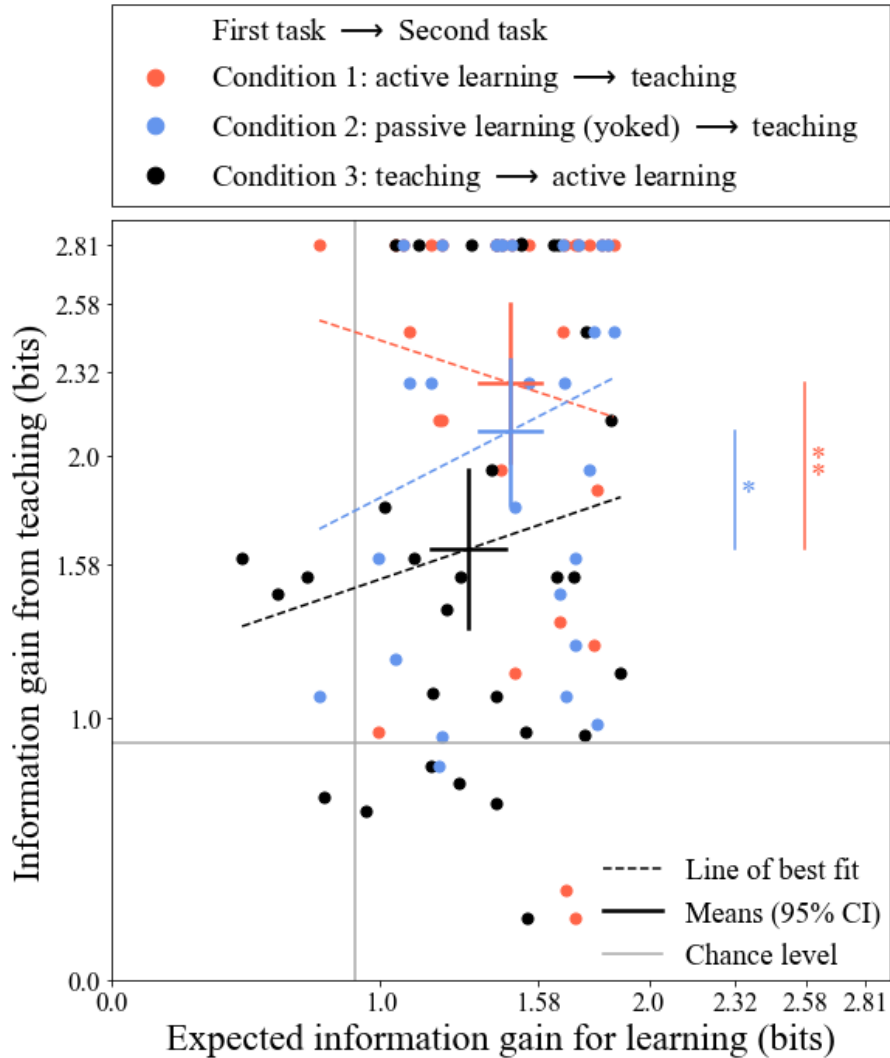


Figure 3.2: Teaching and learning performance across the three conditions. Each dot represents the information gain for one participant, averaged across the three trials of each task. Crosses represent the 95% confidence intervals for the group means. Dotted lines represent the expected mean information gain from teaching as a function of expected information gain. Gray lines mark chance performance (see Supplementary Information for details). The maximum information gain for the task is 2.81 bits. The asterisks mark significance levels in independent t-tests (* $p < .05$, ** $p < .01$).

group differences in learning performance favoring the group that completed the active learning task first, and learning performance is correlated with teaching performance, then the condition effect could be just an artifact. In order to eliminate this possibility, a regression was performed on teaching performance with both the group (active learning before / after teaching⁴), learning performance, and their interaction as predictors. The group difference remained significant, $\beta = .62, p = .01$, when controlling for expected information gain in learning, which was not a significant predictor of teaching ability, $\beta = .08, p = .81$, nor did it interact with the group effect, $\beta = .68, p = .3$. [Figure 3.2](#) shows, for each condition, the estimated (non-significant) slopes for information gain from teaching as predicted by expected information gain for learning. Coupled with the fact that the difference in active learning performance between the two groups was not significant, $t(57) = 1.77, p = .08, BF = 1.28$ in favor of the null hypothesis, this suggests that the effect of the manipulation was not mediated by prior differences in active learning performance. To investigate this issue further, the two groups were repeatedly resampled with replacement such that the learning performance between groups could be matched and fixed at different levels. Comparing the teaching performance across these resampled groups confirmed the advantage of those who completed the learning tasks prior to the teaching task (the 95% CI of the mean of the resampled groups' differences did not include a null effect).

The second prediction of the study was that active learners would gain a larger benefit from learning before teaching than the yoked passive controls. Active learners fared on average only slightly better in the teaching task than their passive learning counterparts who were shown the same labelled data, with an average difference of .18 bits, 95% CI [-.11,.47]. The dyads' performance is illustrated in [Figure 3.3](#). The difference was not significant in a paired t-test, $t(28) = 1.29, p = .21, BF = 2.39$ in favor of the null. It should be noted though that the paired comparison was underpowered (post-hoc power = .24) given the magnitude of the observed effect size. Further, this is possible that the relatively high rates of ceiling performance masked the within dyad effect.

As stated above, there was no significant correlation between learning and teaching per-

⁴The same pattern of results was found for the difference between the group learning passively and then teaching, and the one teaching before active learning.

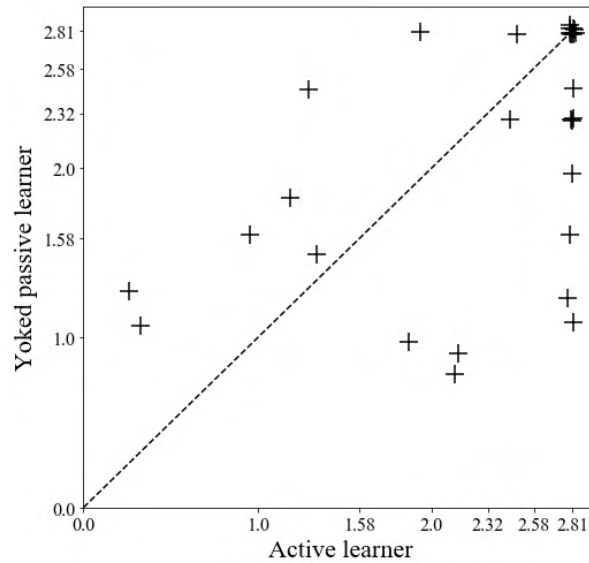


Figure 3.3: Teaching performance for the active-passive learning dyads. Each dot represents the information gain from teaching for one dyad. In dyads situated under the diagonal identity line, the active learner was the better teacher. A small Gaussian scatter was applied to make overlapping dots visible.

formance. However, we observed a trend of increased within-dyad differences in teaching performance for poorer active learning performance (Figure 3.4). This observation remains purely exploratory as this was not a planned analysis and, moreover, the current design lacks power for such a post-hoc test.

Mixed effects analysis

The best-fitting model contained the condition, $F(2,85) = 4.30$, $p = .02$, and trial number, $F(2,174) = 6.93$, $p = .01$, as fixed effects, alongside a participant level random intercept ($SD = .70$). The addition of the random intercept was judged meaningful based on the magnitude of the variance at the participant level ($SD = .70$). It also led to a reduction in BIC, from 796.6 for the fixed effects only model to 749.7.

The previous results regarding the condition effect hold, with a significant estimated difference of .63 bits, $se = .22$, $t(85) = 2.94$, $p = .01$, between the active learning first and teaching first conditions. Similarly, no significant difference was found between active and passive learners, estimated difference of .32 bits, $se = .22$, $t(85) = 2.94$, $p = .01$. Additionally, teaching per-

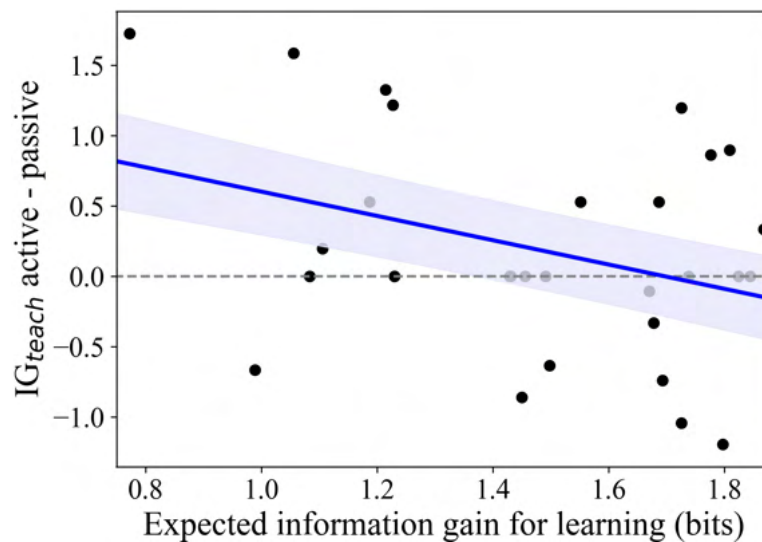


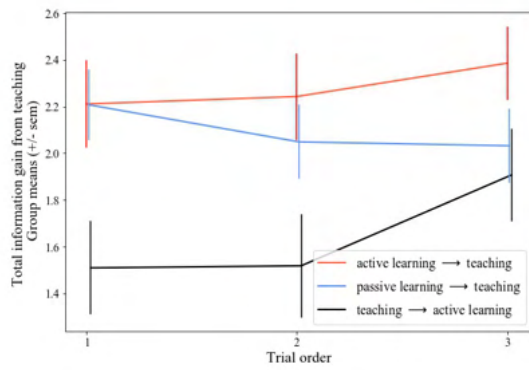
Figure 3.4: The difference in teaching information gain within dyads of active and passive learners as a function of the expected information gain for (active) learning. The fitted OLS regression line is shown alongside its 95% confidence bound. Learning performance was not a significant regressor of the difference in teaching, $\beta = -.86$, $p = .07$. The predicted within-dyad teaching difference was .60 bits, $p=.03$, at a one bit expected learning entropy and decreased to zero for dyads with high expected information gain.

formance improved from the first to the third trial by an estimated .38 bits, $se = .11$, $t(174) = 3.30$, $p = .01$. However, performance improvement from the first to the second trial was not significant, .02 bits, $se = .11$, $t(174) = .17$, $p = .87$.

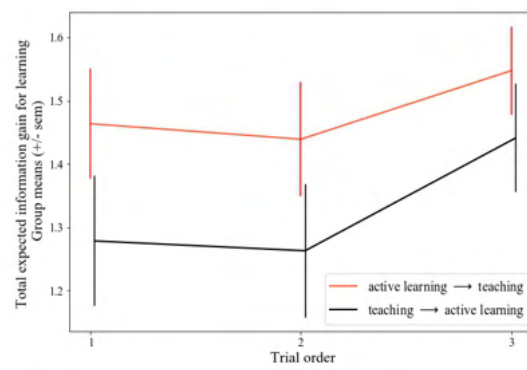
Figure 3.5 illustrates the temporal changes in teaching and learning performance aggregated by condition. While the interaction between the condition and the trial order was not statistically significant, it would appear that teachers who were passive learners first did not improve their performance across trials, through repetition of the task, unlike the teachers from the other two groups.

Decisions about the boundary location

In the active learning first condition, the mean of the best-fit individual r values was .79 (SD = .35), whereas for those completing the active learning following teaching it was lower, .58 (SD = .42). Yoked controls had the smallest average r , .51 (SD = .38). Active learners made better inferences about the boundary location than their matched controls as the average within-dyad



(a) Teaching performance



(b) Active Learning Performance

Figure 3.5: Average group task performance (\pm SEM) broken down by trial number.

difference in estimated probability r was .28, $t(28) = 2.99$, $p = .01$, $BF = 7.29$. The order of the active learning task led to marginally significant differences in an independent t-test, $t(57) = 2$, $p = .05$, $BF = 1.37$ in favour of the alternative.

The difference in r within active-passive learning pairs did not correlate significantly with differences in teaching performance, $r(26) = -.28$, $p = .13$ (see Supplementary Information 3.6.2 for illustrations).

3.3.4 Conclusion

An improvement in teaching performance was observed for participants who engaged in active learning prior to teaching. Three active learning trials, using different stimulus sets than those used for teaching, were sufficient for the majority of participants to gain insight into the optimal solution of the teaching problem on the first attempt. Furthermore, they were able to draw on their experience as learners even though at the time of learning they had not been aware that the teaching task would follow.

The poor performance of participants with no learning experience resonates with previous findings of (Khan et al., 2011), who used a boundary teaching task as well, but did not constrain the example set size by their design. It seems that simply asking teachers to generate the minimally sufficient number of examples for optimal teaching was not enough to solicit the

optimal solution.

The fact that the active learning benefit, relative to teaching first, was not modulated by the initial active learning performance suggests that active learning can improve teaching across the board, for poor and good active learners alike. However, prior active learning performance may play a role in differentiating teachers in a more complex teaching scenarios. Indeed, the surprising lack of a significant correlation between active learning performance and subsequent teaching performance can be explained by ceiling effects.

The impact of passive learning on teaching, relative to the baseline teaching first group, was smaller than that observed for active learning. However, we did not find a significant effect in the matched comparison between active and yoked passive learners. It is important to note here that the current task can be thought of as an insight problem, which means that there was less scope for observing gradual differences in performance. Further, once insight was achieved in the learning task, the solution was easy to verbalize, allowing the optimal strategy to be explicitly transferred to the teaching task.

On the other hand, for poor performing learning dyads, we observed a difference in the predicted direction. This suggests that in a more complex and ecological task in which the learning is more gradual, and the optimal solution is explicitly unknown to participants, active and yoked passive learners are likely to diverge more in terms of teaching performance. This would provide evidence for a more automatic, implicit link between active learning and teaching. In such a future teaching task it would also be interesting to examine whether the differences between active and passive learners, matched for information content, are moderated by the quality of the queries they both observe. Specifically, it should be tested whether the negative linear trend we observed generalizes to non-insight tasks.

Lastly, it is surprising that those who performed the teaching task prior to the active learning task did not differ in their expected information gain in the learning task, and, if anything, performed poorer than their counterparts who started by active learning. This resonates with previous experimental evidence from the developmental literature that has also highlighted more subtle ways in which being taught can hinder learning, for instance by limiting subsequent

exploration (E. Bonawitz et al., 2011b). It is an intriguing idea that, perhaps, not just the experience of being taught, but also teaching itself, can have an effect on exploration. Alternatively, if we assume that the teaching task is more cognitively demanding as it has a meta-cognitive component engaged in reasoning about the learner's knowledge and inference making, results can be explained by the known effect that an easier-to-harder progression of tasks is beneficial for learning, while the opposite order does not provide an appropriate stepping stone for active learning. On the other hand, Yang et al. (2019) argue that active learning can be re-formalized to also include a meta-cognitive aspect, reasoning that is applied reflexively to one's own reasoning.

To conclude, active learning proved to be a reliable intervention to improve teaching performance, but we did not find conclusive quantitative evidence for a benefit of active learning above and beyond passive (yoked) learning. It is important to investigate if the effect of active learning generalizes to more complex and more ecologically valid tasks, or even between more dissimilar learning and teaching tasks. If it does, it will open the way for quantitative inquiries about whether successful teaching benefits from the ability of taking the perspective of an active learner and as such can be improved by prior active learning.

3.4 Experiment 2

3.4.1 Introduction

Experiment 2 was designed to conceptually replicate as well as extend the findings of Experiment 1. As previously, dyads of active and yoked learners were contrasted in their ability to subsequently teach others how to categorize stimuli. However, the classification learning task employed two dimensional perceptual stimuli and was extended in time. These changes were made in order to address several limitations of Experiment 1.

First, the use of a perceptual learning task with continuous-values stimuli makes it less likely for any strategies acquired in the learning phase to be transferred to the teaching task in an explicit manner. Second, the task of Experiment 2 was expected to result in larger interindi-

vidual differences in both learning and teaching performance to enable testing the relationship between active learning and teaching performance. Across two conditions, participants learned different category structures. This allowed for titration of task difficulty to avoid prior ceiling effects.

In the Rule Based condition (RB), the two categories were defined based on a one-dimensional boundary applied to the two dimensional stimuli, making one feature irrelevant to the task. In contrast, both features were relevant for Information Integration (II) categories. Aside from the RB category structure leading to faster learning than the II condition, it has previously argued that learning of the II category structure is procedural or implicit and relies on predecisional integration of features (Maddox & Ashby, 2004), whereas participants test explicit and verbalizable hypotheses to learn RB categories. Active learners have been shown to engage in protracted use (incorrect) unidimensional rules when attempting to learn II categories (Markant & Gureckis, 2014) on this task.

The teaching task conceptually matched that of Experiment 1: selecting a predefined number of examples to an imagined learner. However, teachers had to sample the examples from learned representations that varied in associated uncertainty (corresponding to how well they had learned the categories themselves). Further, teachers designed the stimuli themselves, rather than making forced choices from a predefined array of possible examples. Previous work has found that people are optimal in choosing the most informative question from a given set of possible questions, and yet they do not manage to generate the most informative questions themselves (Rothe et al., 2018). We predict a similar pattern occurs in the case of teaching. In light of the increased difficulty of the teaching task, we will primarily qualitatively compare performance to the predictions of the rational pedagogical model, but also use metric for teaching that relate to the performance of a naïve learner. To that end, we will compute a metric akin to the number of compatible hypotheses consistent with the teaching set (like in Experiment 1) as well as the area inscribed by the teaching sets as a measure of how much of the stimulus space will be familiar to the learner, and how close are the chosen examples to the taught boundary.

Given that the stimuli used were two dimensional, but category structures could be one or two-dimensional, teachers had to choose strategies to emphasize to their learners whether just one or both features were relevant for distinguishing the two categories. Specifically, since active learners in the II condition tend to choose queries consistent with RB category structures, teachers should use examples that signal that both features are involved in the category definition. However, in the RB condition, it is unclear whether teachers will design examples to rule out possible II category structures, either because they failed they realize these kinds of hypotheses are possible or because they believe they learners, much like themselves, will not consider II category boundaries.

Based on performance during the learning task, and subsequent testing of the inferred categorization, it is possible to measure how effective participants were as active learners, their subjective boundary (the hypothesis to be taught) and how uncertain they were in their final subjective classification boundary. As mentioned in the study motivation, the intuition (which matches predictions of the rational pedagogical model) is that teachers who are more uncertain about their own classification boundary, should provide more extreme examples that will avoid potentially misleading the learner.

A last goal was to explore the order of the teaching examples. Teachers were told ahead of time how many examples they could provide to learners, so we expect that they engaged in planning, as opposed to making a decision about the best example at every time step. However, it is still likely that order effects will be found. Active learners on the task are expected to follow a strategy of first exploring the extremes of the stimulus space and then narrowing their queries towards the location of the inferred boundary. Given that teachers know learners will visualize the taught examples sequentially, we expect teachers to follow a corresponding curriculum teaching approach. Thus, unlike previous instances of batch example teaching, manipulating the order of the teaching examples could be highly relevant for successful teaching on this task.

The learning task was closely adapted from Markant and Gureckis (2014), who in turn adapted it from Ashby et al. (2002) for use with active learning using the (Nosofsky, 1989) (circle and radial line) stimuli. The original Markant and Gureckis (2014) study varied the

learning method in four conditions: active learning, yoked passive learning, aware yoked passive learning (learners were told they were yoked to another participant), and reception learning (no-feedback supervised learning). Due to the limited participant pool available for lab-based experiments, only the two conditions directly relevant for the current hypothesis were used, although all four conditions would present interesting additions to Experiment 2.

3.4.2 Methods

Participants

81 participants⁵ (61 female, $M_{\text{age}} = 26.67$ years, range = 18 - 50 years old) were recruited for a lab-based experiment from multiple of participant pools: the university online research participation system, the student union of a local university, first year undergraduate students enrolled in a Psychology course, and personal contacts naive to the purpose of the experiment. As a function of the recruitment platform, participants volunteered their time in return for course credits, received monetary compensation or vouchers worth 1,500 HUF (approximately €4 at the time of participation).

Sample sizes were calculated based on effect sizes estimated from Markant and Gureckis (2014) results. The sample size was reduced from 30 participants/ condition in the original experiment to 20 participants/ condition. This ensured a power of approximately .6 for the main comparison of the categorization performance between active and yoked learners.

Ethical approval for the experiment was obtained from the United Ethical Review Committee for Research in Psychology (EPKEB) in Hungary.

Procedure

All participants completed the learning task first and the teaching task second. Participants were not informed that they will be completing a teaching task before starting the learning

⁵One participant from the active learning condition was excluded from the analysis and replaced as data inspection revealed they did not comply with task instructions and queried the same edge of the stimulus space throughout the entire task.

task. At the end of the learning task, participants were prompted to inform the experimenter, who then explained the teaching task.

The same cover story was used during learning as in Markant and Gureckis (2014). Briefly, participants were told that they would see depictions of loop antennas that received one of two television channels. Which channel was received depended on how the antennas looked and their task was to distinguish Channel 1 receiving antennas from Channel 2 receiving antennas. As a function of the learning condition to which they were assigned, participants were then told that they would learn about antennas either by designing their own and checking which channel they receive, or by being shown antennas together with the channel they received.

Following the instruction presentation, the average experiment duration was 44 minutes for participants assigned to the active learning condition and 21 minutes for participants assigned to the yoked condition.

Active learning

At the start of each active learning trial, a randomly generated stimulus was presented on the screen. 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 ('X' or 'Z'). 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 make manipulate at least one dimension of the stimulus to be able 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 1,500ms.

Yoked passive learning

Each trial started with the brief presentation of a fixation cross (250ms). Then, the stimulus was shown for 250ms 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 screen. Participants could only advance to the next trial if their answer was correct. Responses were elicited from participants to ensure that they attended the stimuli during the otherwise passive learning phase.

Crucially, the stimuli shown to yoked learners were the stimuli designed their paired active learner, and were presented in the exact same order.

Test

Test trials were identical for the two conditions. A stimulus was presented on the screen and the participant had to press one of two buttons corresponding to its category (i.e. whether they thought the “antenna” was more likely to receive Channel 1 or Channel 2). Participants were not provided with trial-by-trial feedback, and were only shown the percentage of correct answers at the end of every block. Unlike in Markant and Gureckis (2014), participants were not asked to provide a confidence judgement after every categorization choice. This decision was made in order to shorten the length of the experiment (which was increased by the addition of the teaching phase), given the fact that it was not directly relevant for the current hypothesis.

Test stimuli were sampled uniformly for the quadrants of the stimulus space to ensure that the test phase did not bias participants’ representation of the two categories and to ensure that across all conditions chance performance was 50%.

Teaching task

The teaching task required participants to design stimuli using the same procedure that active learners had previously used to generate queries. However, this procedure was new to the yoked learners. In order to ensure that passive learners had a comparable mastery of how to produce the examples, participants were allowed to first explore the setup for as long as they wanted and then they performed an additional practice task. Participants were shown two antennas on the screen in two different colors: a fixed target blue antenna and a randomly generated black antenna that they could modify. Their task was to manipulate the black antenna until it perfectly matched the target antenna. All participants performed at least 4 trials of the practice task in the presence of the experimenter. Participants were given additional practice trials if they were not sufficiently competent at designing the antennas.

Participants were told they are going to teach another participant who is yet to take part in our experiment which antennas receive Channel 1 and Channel 2. It was stressed that these participants were naïve about the stimuli and that they too will be performing a categorization

test like the one that the participant just completed. Participants were then told that the only constraint for the teaching task was that they can only provide a maximum of 6 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. 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.

Following the example selection, participants were prompted to type answers to an open-ended questionnaire about their teaching strategies, if any, and how they would have taught another participant if they could verbally communicate to them.

Design

The learning task was closely adapted from Markant and Gureckis (2014). A 2 (learning type) by 2 (category structure) between participants design was used. Participants were randomly assigned to one of the two category structures (RB or II). Each yoked participant was paired to an active learner.

All features governing the presentation of the stimuli were counterbalanced: the feature relevant for classification in condition RB (orientation or circle radius), the diagonal (first or second) used for the II categorization, the mapping between the two keyboard buttons and the features, the mapping between the mouse direction of movement and the relative direction of changes in the stimuli presented on the screen.

The learning tasks consisted of eight blocks. Each block contained 16 training trials followed by 32 test trials.

The number of teaching examples was constrained to a predetermined value to ensure a fair comparison between participants on the teaching task. A fairly low number of teaching examples was preferred in order to both stress to participants the importance of selecting good

examples and to result in well powered comparisons against predictions based on random sampling (as pedagogical sampling and random sampling become indistinguishable with larger numbers of samples). Further, 6 examples would be sufficient to optimally describe the two categories. The exact number of teaching examples was determined based on a small pilot (5 participants⁶ who were learners in the RB condition) using the same active task and in which participants were allowed to choose up to 12 teaching examples. The pilot was primarily used to check if the active learning stimulus selection procedure was implemented correctly and that participants found it easy to use.

Stimuli

The stimuli were circles defined by the size of the radius and the orientation of the diameter. These stimuli are widely studied in the categorization literature. The two features that define the stimuli have been shown to be perceived as independent by Nosofsky (1989).

Minimum, maximum and range values of the radius of the circles were determined in relation to the computer screen size (27inch) and sitting distance. The same range was used for every participant, but the minimum value was pseudo-randomly sampled given the maximum possible stimulus constraint. For the orientation, a range of 90 degrees was used. This avoided the use of circular variables (150 degrees in the original design), and the value of the minimal angle was pseudo-randomized for every participant.

The deterministic boundary between the two categories in the rule based condition was always midway through the stimulus space and in the II condition it was either the first or the second diagonal. Since every participant had a different boundary in perceptual space, we rescale stimuli to an abstract stimulus space for all analyses and illustrations (although arbitrary, the Markant and Gureckis (2014) ranges are used for ease of comparison).

The task was coded in *PsychoPy*³ and the anonymized data are available at tinyurl.com/27em6ajf.

⁶At the time of the pilot, we were not aware of the Avrahami et al. (1997) experiment. They found that participants chose on average 7 examples, confirming with a much larger sample that this is a range comparable to what teachers would have done naturally.

Data analysis for the learning task

The main data analysis used by Markant and Gureckis (2014) was replicated for the learning task. Active and yoked participants were compared in terms of their test performance across blocks⁷. We show both ANOVA between group comparisons as in Markant and Gureckis (2014), as well as within-dyad comparisons for categorization accuracy.

The performance of the active learners was quantified by the distance from their queries to the true boundary, as well as to their subjective boundaries. Individual subjective boundaries were fitted for every categorization test block to allow tracking of how participant's boundaries changed with learning and to provide a more suitable metric for the sampling behaviour of active learners. This is especially relevant in the II conditions where a lot of participants did not converge to the true boundary by the end of the experiment.

The subjective boundaries were fitted 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 3.5 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 is the determinism of the boundary.

$$P(x^{trial} = CH1 | \theta, b, \sigma) = \frac{1}{1 + \exp(-\sigma(x_1^{trial} \cdot \cos(\theta) + x_2^{trial} \cdot \sin(\theta) - b))} \quad (3.5)$$

The interpretation of the fitted parameters (θ, σ, b) is visually illustrated in Supplementary Information Figure 3.31. For every decision-bound model, a random-response model was fitted according to which participants choose a category according to a given probability without using the location of the stimulus in the stimulus space. Comparisons were performed between the random and decision-bound models using the Aikake Information Criterion (AIC). Figure 3.35

⁷Response times were collected for trials/decisions in all tasks, but are not presented in the data analysis.

in the Supplementary Information shows the AIC values for the random and decision-boundary model for every participant. Participants who were not better fit by the decision-bound model were eliminated from the analysis.

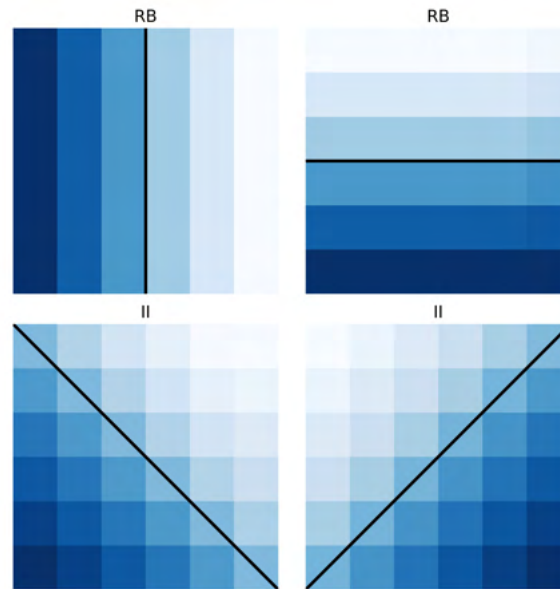
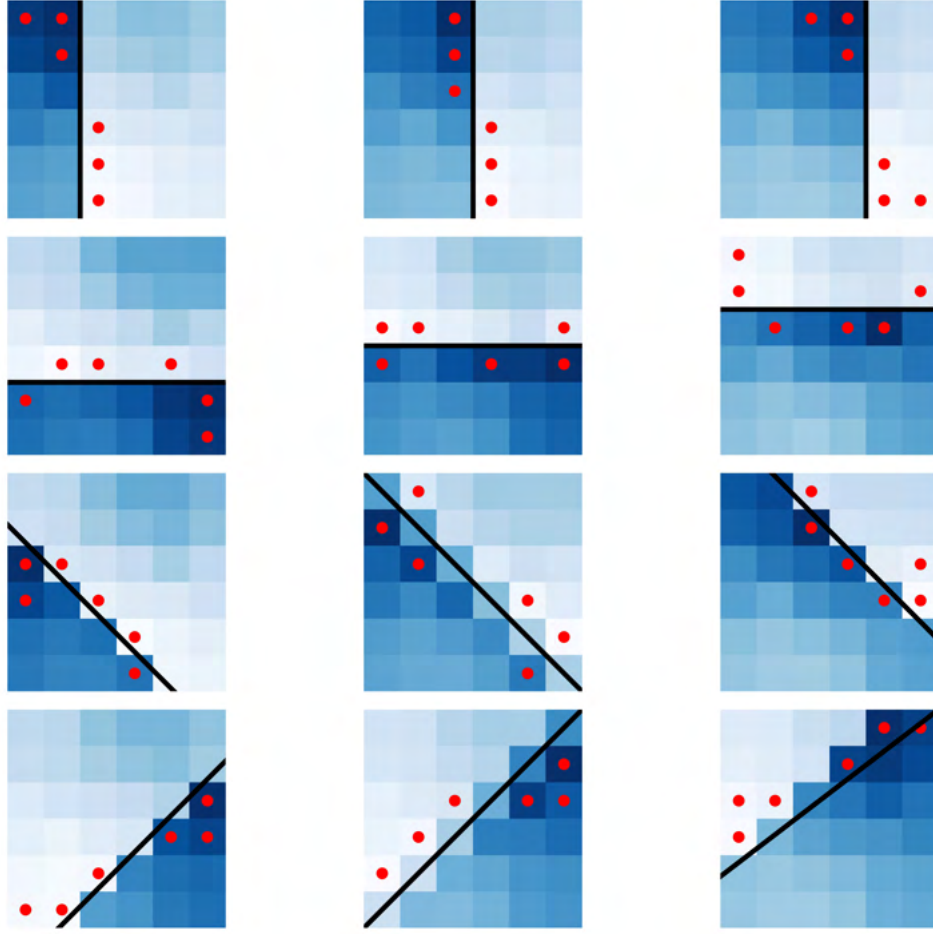


Figure 3.6: Random sampling of stimuli to offer as examples for the four boundaries used in the task. The farther from the boundary an example is, the more likely to be chosen as a category example.

Teaching task quantification and predictions

As a first check, we test whether the samples pooled across participants in the different conditions significantly departed from random uniform sampling across the stimulus space. Specifically, the empirically observed distribution of samples in the stimulus space was compared to uniform sampling using the Kolmogorov-Smirnoff test (KS). The KS test computes the distance between two empirical cumulative distributions of two samples to assess whether the samples likely came from the same distribution.

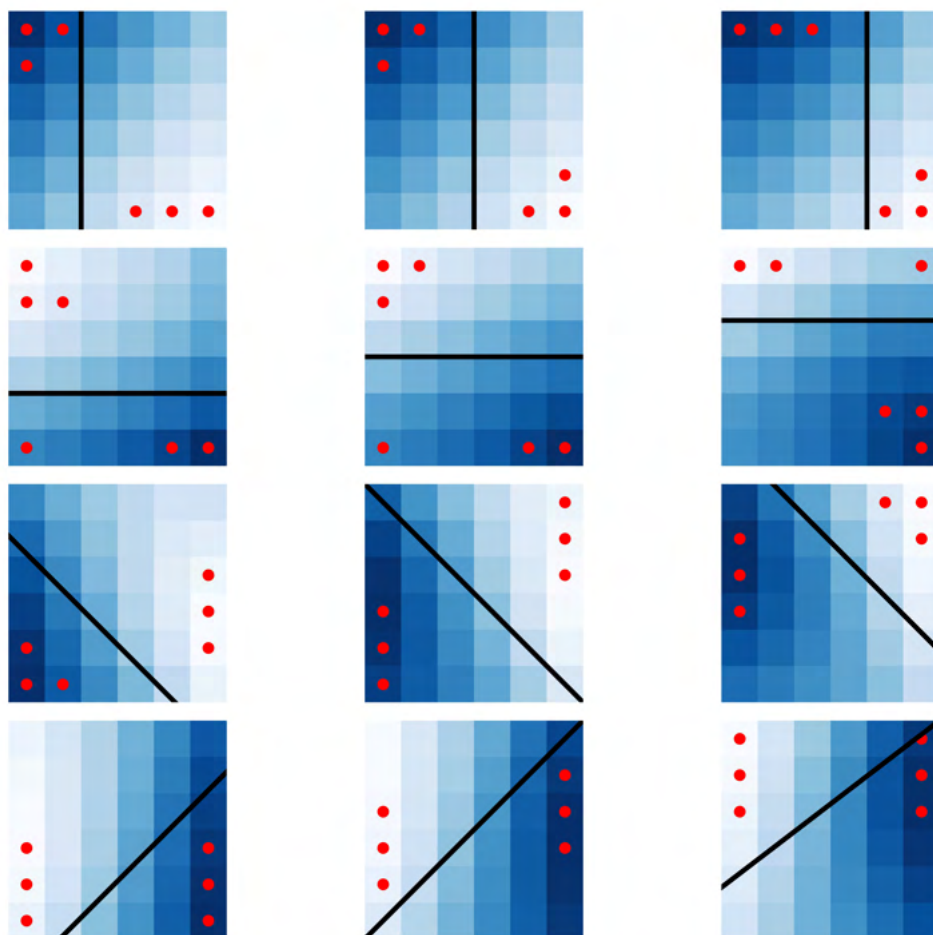
Given that the participants' behavior in the categorization test is well captured by the decision-bound model, it can be used to compute the likelihood of examples under random sampling. This intuitively predicts that an example is more likely to be selected if its orthogonal distance to the category boundary is larger (see [Figure 3.6](#)). This leads to a large number of



(a) Large σ - Boundary is highly deterministic

equiprobable teaching tests.

Predictions were also generated from an iterative pedagogical sampling model (Shafto et al., 2014). To make the computation of the likelihood and posterior tractable, the stimulus space was discretized into a 6x6 grid, and the probability of a stimulus being selected as a teaching example was calculated only for the resulting 36 locations. The set of possible hypotheses considered was the set of lines defined by combinations of θ values of $0^\circ, 45^\circ, 90^\circ, 135^\circ$ (corresponding to the orientations of the true boundaries used in the task) and center biases of -1, 1, and 0. This resulted in 12 possible hypotheses. The teaching sets were all the possible (order invariant) sets of six examples with all permutations of category labels such that three examples belong to one category and three to the other ($C_6^{36} \cdot \frac{6!}{3!3!}$ example sets). The α parameter was set to 1 and the starting probability of choosing an example was proportional to the factorized



(b) Small σ - Responses are stochastic around the boundary

Figure 3.7: Predictions from the pedagogical model for the likelihood of choosing examples at different locations in the stimulus space. The most probable example sets are overlaid in red and the hypotheses are drawn in black.

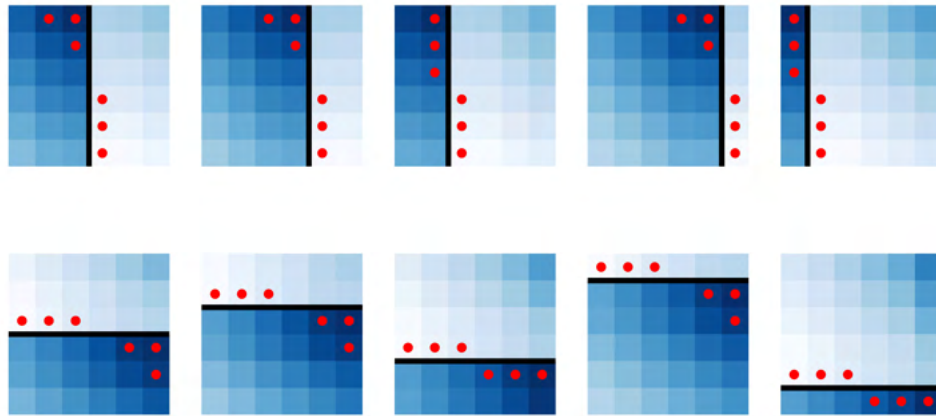


Figure 3.8: Predictions of the pedagogical model for teaching only RB categories. Red dots represent the most likely example set under pedagogical sampling.

probability of the stimuli belonging to a given category.

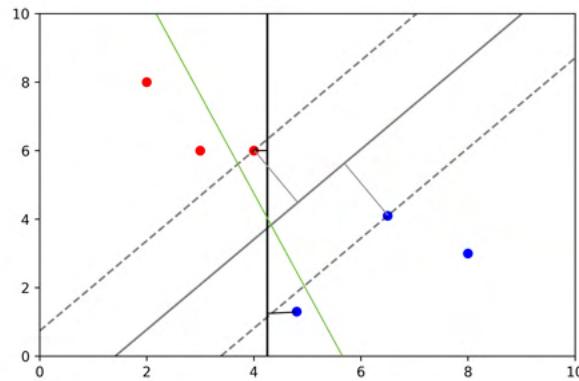
Given the fact that each participant had a different subjective boundary and associated boundary determinism, we tried to extrapolate general predictions from the pedagogical model to test on the current sample rather than perform a direct model fit to the behavioral data.⁸

Figure 3.7 shows the predictions for a case in which high and low values of σ . When the boundary is highly deterministic, the most likely pedagogical examples are situated in close proximity of the boundary. Note that these examples do not necessarily uniquely define the hypothesized boundary. For instance, in Figure 3.7, most likely examples for the vertical boundary running through the middle of the stimulus space are also compatible with a horizontal boundary. However, examples vary maximally in the categorization irrelevant dimension, and vary minimally in the relevant dimension. In this case, the use of very low or very high variability is intended to highlight the saliency of features. In order to explore whether participants used this strategy, we computed for every participant the ratio of variance in stimulus distances from the categorization dimension to the variance in distances to the categorization irrelevant di-

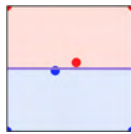
⁸We did calculate the probability of each participants' teaching set based on a strong sampling model on the discretized 6x6 grid and using their individual inferred parameters. We found the sets for most participants to be highly consistent with strong sampling. Unfortunately, it was not possible to do the same for the pedagogical sampling assumptions to then apply a LRT due to computational demands.

mension (orthogonal to the boundary). Another intuitive pattern for examples compatible with multiple boundaries is that the to-be-taught boundary is the ones that ensures minimal separation between the examples from the two categories (Figure 3.7 subfigure on row one, columns 1 and 3). In order to check for this in the behavioral data, we computed the orthogonal distance between examples closest to the boundary on opposite categories (detailed below).

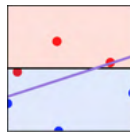
Lastly, predictions were made for teaching RB categories assuming a strong prior for axis-orthogonal boundaries (see Figure 3.8) for comparison with the teaching samples for the RB category structure. Here all possible vertical and horizontal boundaries were used as hypotheses (10 boundaries on the 6x6 grid). The pattern of results was the same as described above.



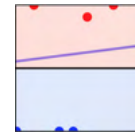
(a) Example set that demonstrates the possible strategies of a fictitious learner. There are multiple boundaries that perfectly separate the examples (but the green line does not). The black line minimizes the orthogonal distance between the 2 closest examples. The gray line is the maximum-margin distance boundary which maximizes the orthogonal distance between the two closest examples of different class. If these samples would be randomly generated, a learner would not have any reason to prefer one boundary over another (absent any priors about the category structure. Learners who receive pedagogically sampled examples may believe that the black boundary is more likely because it minimizes the distance between classes.



(b) Participant using the maximum area strategy.



(c) Participant minimizing the set of consistent boundaries.



(d) Participant who scores low on both metrics.

Figure 3.9: Teaching samples (dots in the two dimensional stimulus space) offered by two experiment participants in the RB condition. Purple lines represent their subjective boundary inferred based on the last test block. Dot colors represent the category labels chosen by the participant and area colors represent the ground truth.

In addition to predictions that take into account learners who make pedagogical sampling

assumptions, we quantified teaching performance with reference to naïve learners. Thus, for every teacher's set of samples, we determined the subset of linear boundaries that were compatible with the spread of examples in the stimulus space. The reasoning is that a good set of examples should result in only a limited set of probably hypotheses (with the hypothesis to be taught being among them), and thus decreasing a naïve learner's uncertainty about the correct boundary. A boundary was deemed consistent with an experiment set if it perfectly (linearly) separated examples from the two conditions, and the resulting labeling of items left or right of the boundary matched the actual category choices. The space of possible boundaries was determined by the bivariate space of all θ and b values such that the line they defined crossed the stimulus square. The number of boundaries consistent with an example was then divided by the total number of boundaries considered (a fine uniform grid over the boundary space).

This metric assumed that learners would treat the task as a linear classification task. Although this seems like a likely assumption, this may not necessarily be the case when receiving such a small number of examples, and learners might not generalize the two categories beyond the examples they had been shown. In this case, to cover the entire variety of exemplars contained within a category, an intuitive strategy is to show learners the examples that inscribe the largest possible stimulus area. This strategy has considerable overlap with the minimum consistent set strategy described above, but the two can be dissociated. [Figure 3.9](#) shows individual participant data that illustrates these strategies. While the orthogonal distances of the examples to the subjective boundaries in both cases are very small, a naïve participant seeing the labeled examples in (b) might draw a diagonal boundary leading them to wrongly classify a large area of the stimulus space. However, in panel (c) all consistent boundaries are very similar and the potential for misclassification is low.

The area of the polygons⁹ inscribed by the examples given by the participant for each category was calculated separately. The polygon areas corresponding to each category were summed up after removing the intersecting areas. The larger the area provided by the teacher,

⁹The majority of participants chose to give three examples for each category, meaning that a triangle is defined for each half of the stimulus space. For the few participants who decided to use an asymmetric number of examples across the two categories, we could not calculate an area for one of the categories, so only the area corresponding to examples from one category is used.

the more uncertainty is removed for a potential learner about how the stimulus space is divided into the two categories, under the assumption that learners believe there is continuity in the feature space occupied by a category. This metric is an analogue in perceptual space for the one used by Shafto et al. (2014) in the rectangle game. As before, this metric can be compared with the distribution of areas produced under random sampling of examples from the two categories.

Determining which boundaries are and are not compatible with an imagined example set is in all likelihood a very taxing strategy for teachers. It is a difficult task not just because of the infinite hypothesis space, but also because it requires anticipating that another participant (without a similar learning experience) may form hypotheses about the boundary consistent with a different category structure. To measure this, the ‘boundary distance’ was computed as the minimal orthogonal distance of the closest examples on either side of the subjective boundary of the teacher.¹⁰¹¹ Moreover, this measure captures the intuition behind the pedagogically generated examples, that learners who view examples generated pedagogically will choose a boundary that minimizes the separation of the examples from the two categories. Figure 3.9a provides an illustration of the learner’s inference problem.

It can be expected that poor learners, who were still uncertain about the location of the boundary at the end of the training, would choose examples fairly far apart to avoid mislabeling them, and therefore, misguiding the learner. To that end, if learners are aware of their own limitations as learners, there should be a negative correlation between the σ parameter fitted in the last test block of the experiment and the boundary distance measure described above (as well as average distance from boundary). One concern regarding this metric is that since it inherently relies on the subjective boundaries of the participants, noise due to not just the participant’s inherent uncertainty in the boundary influences the distances, but also the estimation noise. Thus, it is possible that this measure is not very robust.

Lastly, as the pedagogical model predicts, examples vary considerably along on a dimension parallel with the decision boundary while keeping the distance to the boundary relatively fixed.

¹⁰For one-dimensional stimuli, the distance metric is the continuous equivalent of the number of consistent boundaries (hypotheses) used in Experiment 1.

¹¹This approach is in direct contrast with the canonical solution for classifying linearly separable datasets using SVMs, namely choosing the maximal separation boundary.

Therefore, we checked the distribution of distances from the boundary, but also from the line orthogonal to the boundary to check if participants were using variability as a cue to the relevant classification dimension.

The main hypothesis of the experiment concerned significant differences in the pedagogical sampling efficacy of the active and yoked learners. Paired t-tests were conducted within dyads on the metrics described above. The subjective teacher boundaries, where used, were the ones inferred from the last teaching block. Since it would not be meaningful to investigate the teaching behaviour of participants who did not learn any boundary by the end of the task, participants whose test responses in the last block were better fitted by the random response model were excluded from the experiment.

Further, the data will be visually inspected for any example order effects consistent with curriculum teaching (examples initially further away from the boundary). To test whether there were any consistent strategies within participants, examples will be labelled as extreme, central or close to the boundary and the distribution of example types (in order) across participants will be compared to chance.

Lastly, differences were expected in the final categorization performance of active and yoked learners. This raises the concern that any within-dyad differences in teaching performance stem solely from yoked learners have a more uncertain/poor representation of the two categories. In order to check for differences independent of the accuracy at test, groups of active and passive learners were subsampled with replacement such that they had roughly equal performance at test, and compared in their teaching performance.

3.4.3 Results

Categorization Performance

Overall, results were highly consistent with the findings of Markant and Gureckis (2014). Participants in the RB condition outperformed participants in the II condition at test (see [Figure 3.10](#)). Further, active learners, regardless of condition, were more likely to be correct in

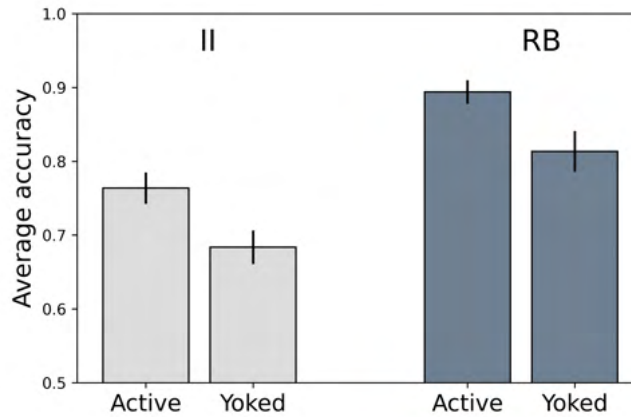


Figure 3.10: Overall accuracy in categorization. Bars represent standard error of the mean.

the categorization test than yoked learners. A 2x2 between participants ANOVA resulted in significant main effects of learning mode, $F(1, 77) = 14.40, p < .001$, and category structure, $F(1, 77) = 37.65, p < .001$, without a significant interaction. These differences remained significant in the last test block of the experiment.

The overall categorization accuracy of active learners was directly compared with that of the passive learners they were paired with in Figure 3.11. The vast majority of active learners surpassed their paired yoked learner. Within-dyad differences 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 3.11 shows how average categorization performance changes across the eight experimental blocks and Figure 3.32 illustrates each individual's learning trajectory. The performance of RB active learners reached almost ceiling levels half way throughout the task, while II active learners barely surpassed 80% accuracy. The performance of passive learners was consistently lower and plateaued earlier. Note that applying a one-dimensional rule in the II condition would result in around 75% accuracy.

The subjective boundaries fit to every test block are shown in Figure 3.12. 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

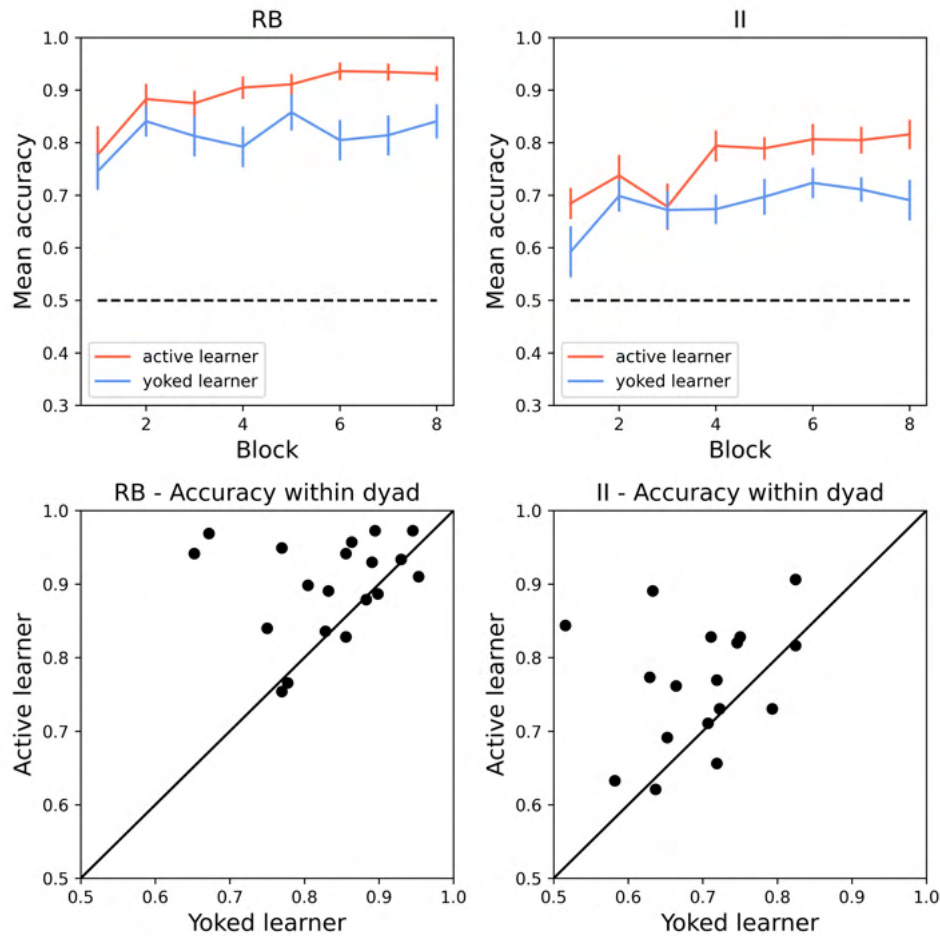


Figure 3.11: Average categorization accuracy by experimental condition and block order (Upper panel). Vertical lines represent the standard error of the mean. Using a unidimensional rule in the II condition would result in about 75% average accuracy. Lower panel shows within dyad categorization performance differences. Each dot represents a dyad. All axes display the proportion of correct responses.

was no discernible convergence pattern by the end of the task, but a large proportion of active learners in the II condition were fitted by boundaries that were not axis-aligned and relatively close to the true boundary. While overall the participants' categorization was better fitted by the decision-bound model, there were participants whose behaviour was better described by the random response model, when adjusting for the number of parameters used by the model. This differed across the two category structures as only 84.37% of test blocks in the II condition were better fitted by the decision-bound model than a random response model (according to the AIC criterion), compared to 93.44% in the RB condition. Further, the final test block responses of 10 (out of 20) yoked learners in the II condition were fit equally well by strict one-

dimensional boundaries, while the same was true for only 3 (out of 20) of the active learners in the II condition. Supplementary Information 3.33 shows individual boundary fits overlaid on participants' choices.

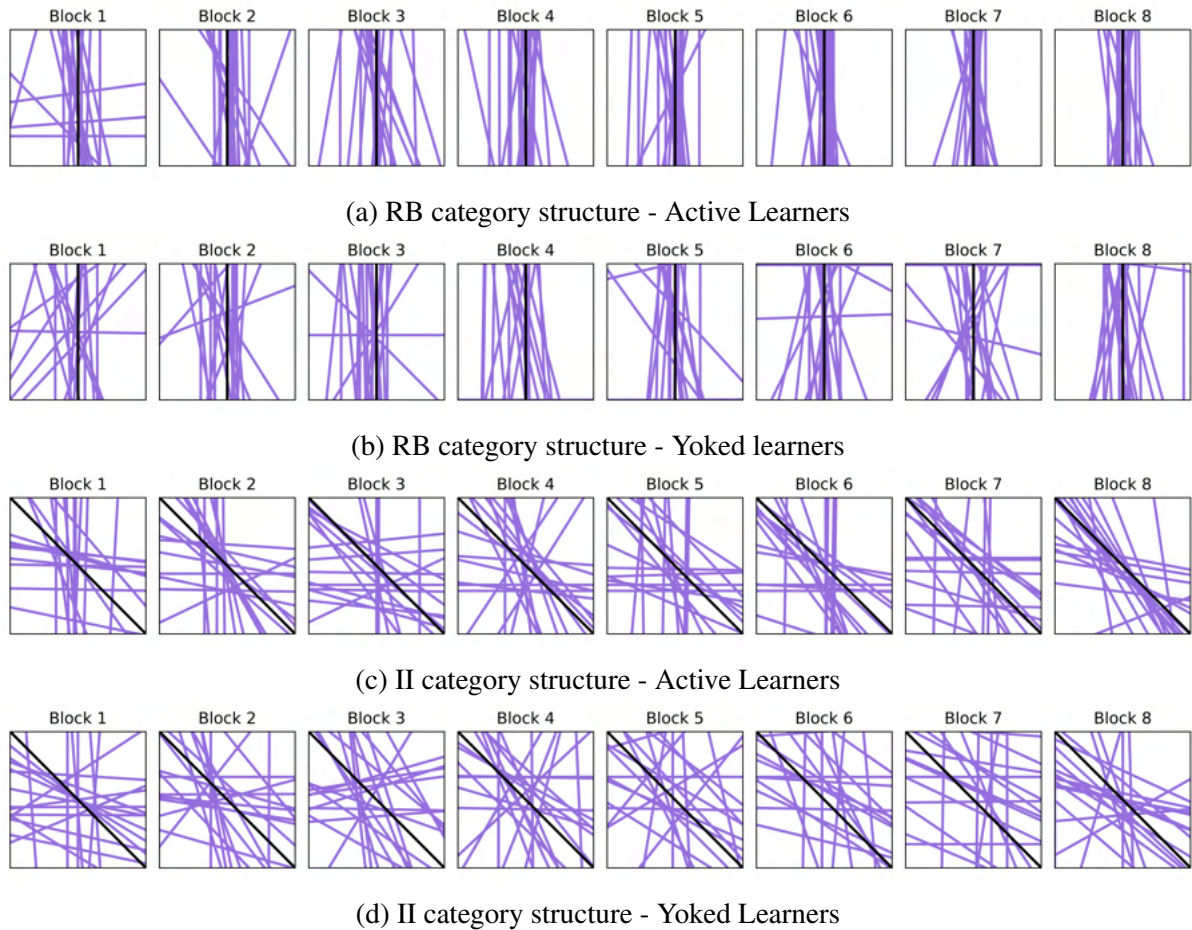


Figure 3.12: Fitted boundaries for all participants (purple) against true boundaries (black). Stimulus spaces were rotated in order to be able to overlay boundaries of all participants.

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. This pattern is consistent with an early exploration of the extremes of the stimulus space. The oversampling of extremes can be seen in Figure 3.14 which shows all the stimuli designed by participants across the eight learning blocks (see individual plots of the queries in Supplementary Information 3.34). Figure 3.13 illustrates the gradual decrease in the average distance of queries to the boundary,

for both the RB and the II condition. 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. 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, $t(19) = -1.56, p = .13, BF_{null} = 1.52$.

There was a strong relationship between the average query distance and the average accuracy at test for active learners (Figure 3.15). 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 : $r(18) = -.20, p = .39$). This lends support to the idea that the quality of the observed data did not impact the performance of yoked learners, confirming Markant and Gureckis (2014) predictions.

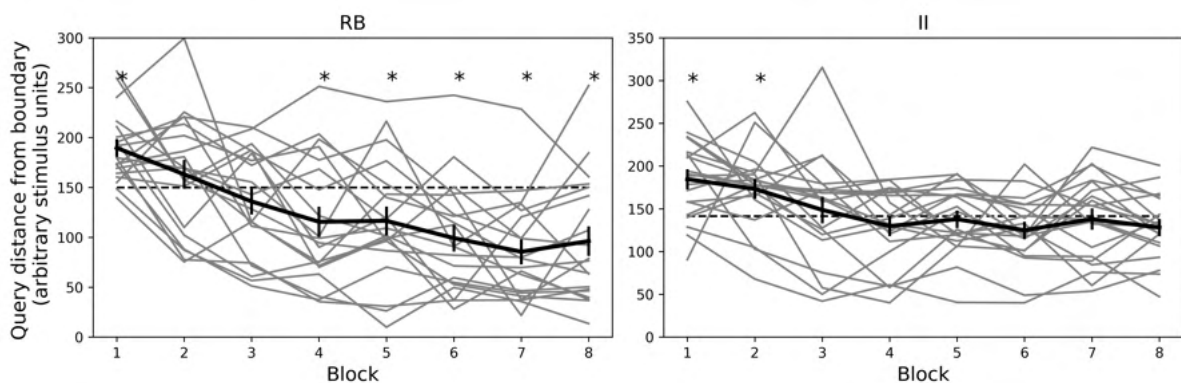
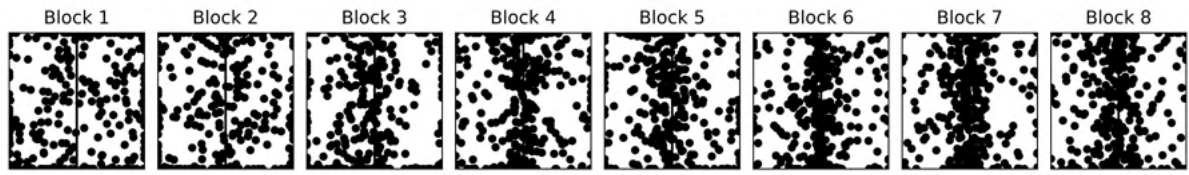
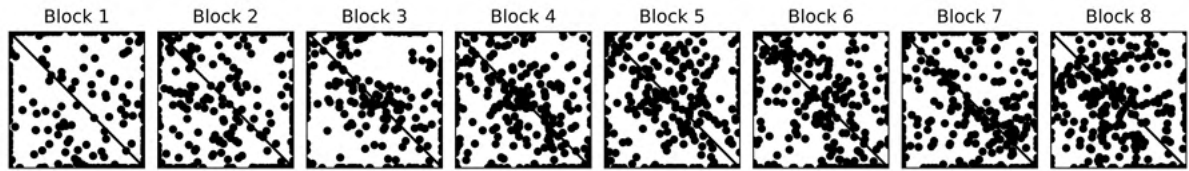


Figure 3.13: Average query distance to boundary for the group (black) and individual participants (gray). Bars represent standard errors of the mean. The dashed line is the expected query distance under random sampling for each corresponding category structure. Asterisks are displayed above blocks where the average query distance differed from the chance significantly in a one-sample two-tailed t-test ($\alpha < .05$).



(a) RB category structure



(b) II category structure

Figure 3.14: Queries made by all participants across the 8 active learning blocks. Each dot corresponds to the angle and radius size that defined a stimulus. Stimuli have been rotated such that the true boundaries are aligned for all participants.

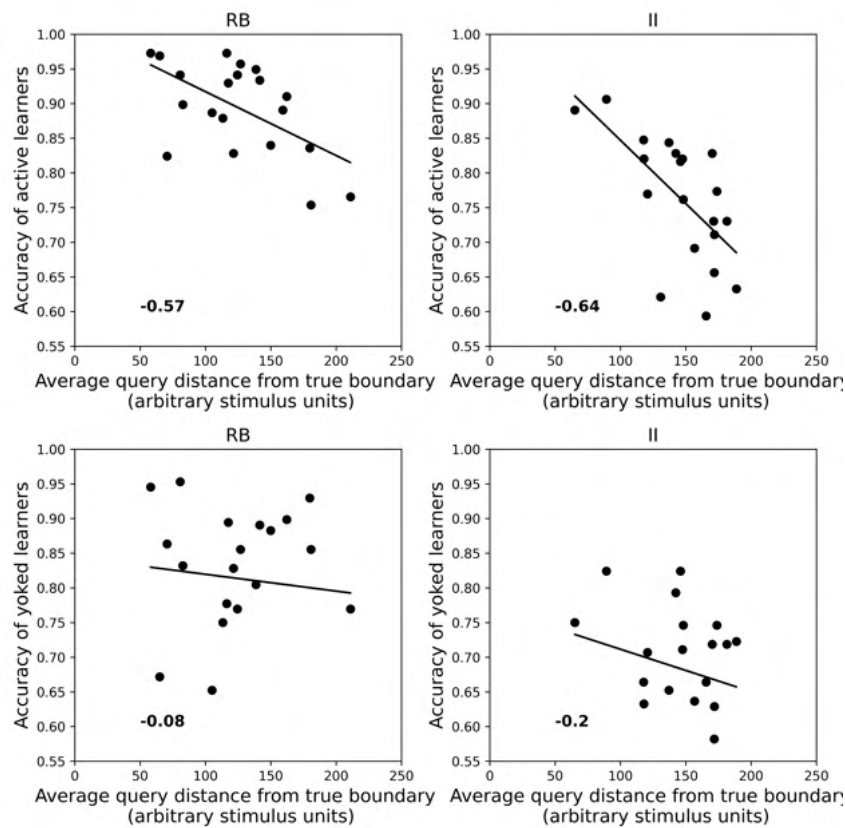


Figure 3.15: Scatterplot of the average query distance from the true boundary against overall categorization performance (of active learners or paired yoked learners). Each dot is a participant. Correlations are significant for active learners (upper plots), but not between an active learner's categorization accuracy and their paired active learner's query distance (lower plots).

Teaching Performance

Participants had high levels of accuracy when offering examples, that is, the examples they gave for a particular category actually belonged to that category 90% of the time. Figure 3.37 shows all examples offered alongside the chosen categories. Participants overwhelmingly chose to give an equal number of examples from the two categories (77.50%). Supplementary Information 3.38 plots the teaching examples separately for each participant.

Distribution of examples (group level)

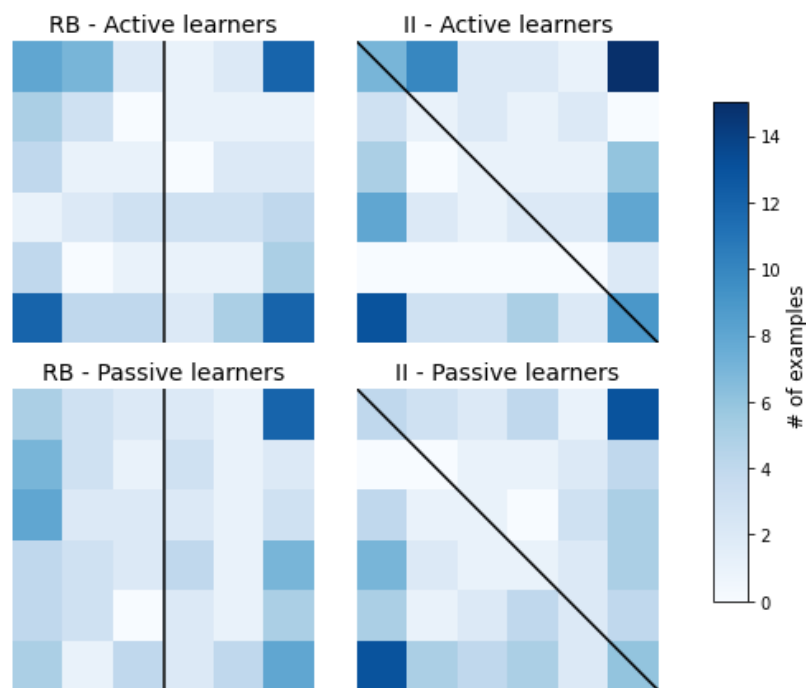


Figure 3.16: Frequency of teaching examples across the stimulus space (pooled across all participants). Color intensity corresponds to the number of examples in each bin. Each cell contains about 3% of samples under uniform sampling of the stimulus space.

Figure 3.16 shows the spread of all teaching samples across the stimulus space. Qualitatively, at the group level, example selection is compatible with pedagogical sampling assuming very low boundary determinism, which predicts oversampling of extreme stimulus values. This was expected, at least for the II category structure, given the variability in boundaries intended for teaching, but was surprising for the RB category where participants were generally very precise in the last categorization test. The distribution of examples under both category structures

was different from uniform sampling. However, for RB categories, the variation along the classification irrelevant feature was indistinguishable from uniform sampling for yoked learners. [Figure 3.6.3](#) details the statistical test results.

Visual inspection suggested that participants oversampled the corners of the stimulus space. This can also be seen in [Figure 3.17](#) which presents distances from the categorization boundary (relevant dimension) and distances from the line orthogonal to the categorization boundary (irrelevant dimension). Maximal distances from the boundaries were over-represented for both category structures. In the II condition, the distribution of examples was at least bimodal (Hartigans' dip test for multimodality: $D = 0.04, p = 0.02$). This was not the case in the RB condition, as there was one clear peak corresponding to maximal distances.

Further, as predicted based on the pedagogical sampling model that participants would choose values that vary more along the dimension orthogonal to the categorization boundary. At the sample level, teachers in the RB condition who were active learners oversampled both extremes of the stimulus space when choosing the value of the irrelevant feature ([Figure 3.17](#), $D = .085, p < .001$). This pattern was consistent with flagging corners of the stimulus space. However, as a group, teachers who were previously yoked learners uniformly sampled values for the irrelevant feature ($D = .04, p = 0.13$). The same applied for II category teachers, both active and yoked.

Differences between active and yoked learners

The primary metric for the quality of teaching was the proportion of boundaries compatible with the examples, the smaller the better. As seen in [Figure 3.19](#), active learners surpassed passive learners in the II condition, $t(15) = -2.26, p = .04, BF_{alt} = 1.82$, but not in the RB condition, $t(15) = -.72, p = .48, BF_{null} = 3.12$. Four participants (2 yoked and 2 active learners) who produced example sets that were not linearly separable, and they were excluded from all the paired tests presented along with their dyad partners. There was no difference in the number of consistent boundaries between the two category structures, $t(62) = -.36, p = .72, BF_{null} = 3.70$.

For every participant, the area of the polygons inscribed by examples from each category

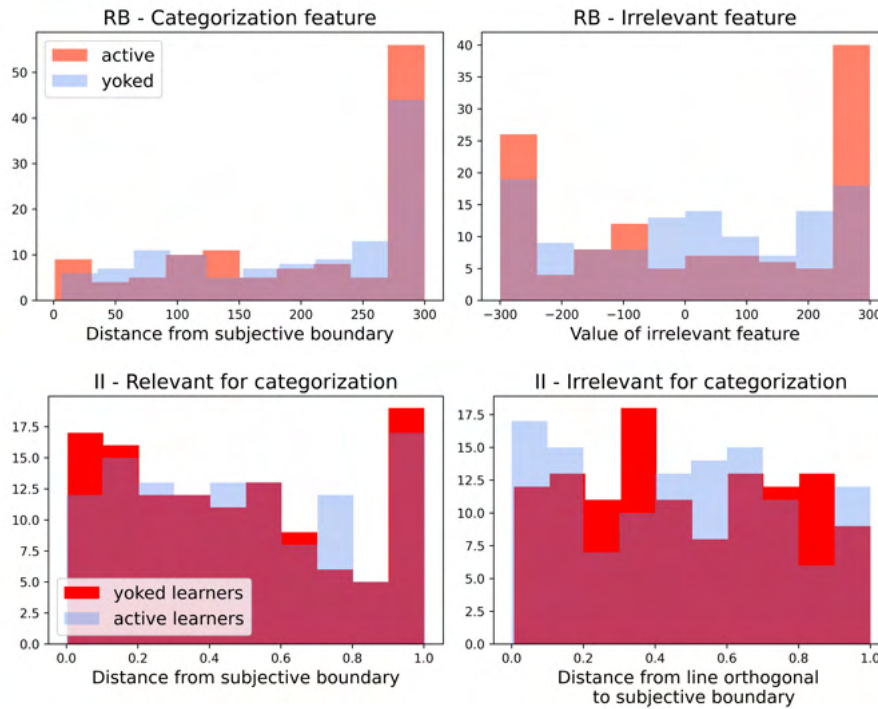


Figure 3.17: Left: Histogram of the orthogonal distances between the chosen teaching examples and the boundary. Right: Histogram of the distances between teaching samples and a line orthogonal to the boundary (corresponding to the irrelevant feature in the RB condition or the opposite diagonal in the II category). The values presented were rescaled to maximum possible distance from the subjective boundary.

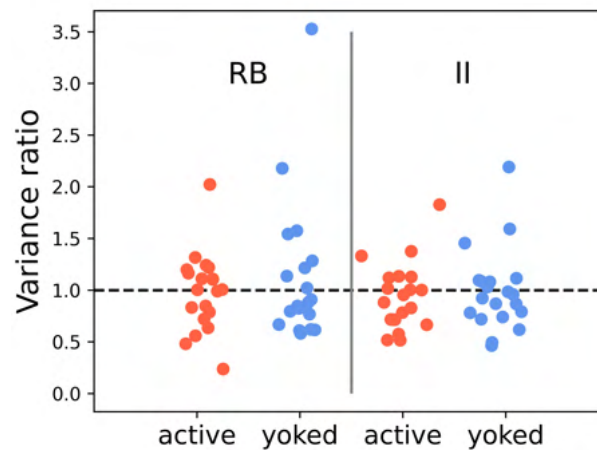


Figure 3.18: Ratio of standard deviations. Numerator: orthogonal distances of teaching examples to the subjective categorization boundary. Denominator: orthogonal distances of teaching examples to the line orthogonal to the subjective categorization boundary. Pedagogical prediction is that the ratio should be smaller than 1.

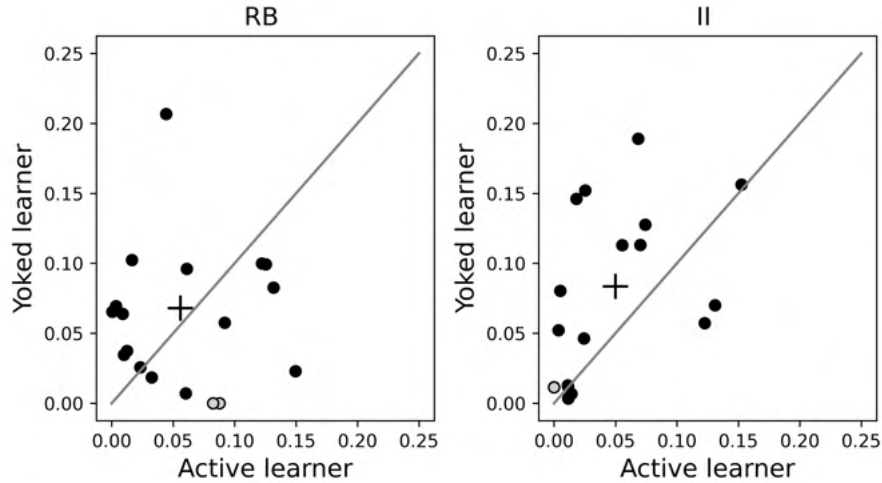


Figure 3.19: Proportion of boundaries compatible with the teaching examples. Each dot is a dyad. Crosses mark the sample means. Grey dots correspond to dyads in which one of the participants did not choose a linearly separable example set and were not included in the analysis.

was calculated, subtracting any potential overlaps between the areas corresponding to the two categories. Overwhelmingly, participants chose three examples in each category, so this corresponded to the area of two triangles. Intersections occurred for only 10% of participants and were generally small. Active learners outperformed their yoked dyad counterparts according to the area teaching metric (see Figure 3.20), selecting examples that inscribed a significantly larger area of the stimulus space in the II condition, $t(17) = 3.66, p < .001, BF_{alt} = 21.13$, but not in the RB condition, $t(17) = 1.87, p = .08, BF_{alt} = 1.02$.

Figure 3.20 shows the distribution of areas that can be expected from random sampling from the two categories calculated based on large simulations ($n = 10^6$). The expected mean area of a random triangle inscribed in a unit square and the area distribution function can also be computed analytically ($\mu \approx .072$, Weisstein (n.d.)). The mean areas produced by participants were larger than the the mean expected by random sampling for active learners, but was roughly the same for yoked learners. For each condition, a two-sample KS test was run to test if the distribution of areas produced by participants significantly differed from the random sampling (simulated) distribution. The KS tests in the active groups found significant differences, RB: $D = .44, p = .001$, II: $D = .56, p < .001$. However, the null was not rejected for the yoked

learner groups: RB: $D = .17, p = .62$, II: $D = .22, p = .31$.

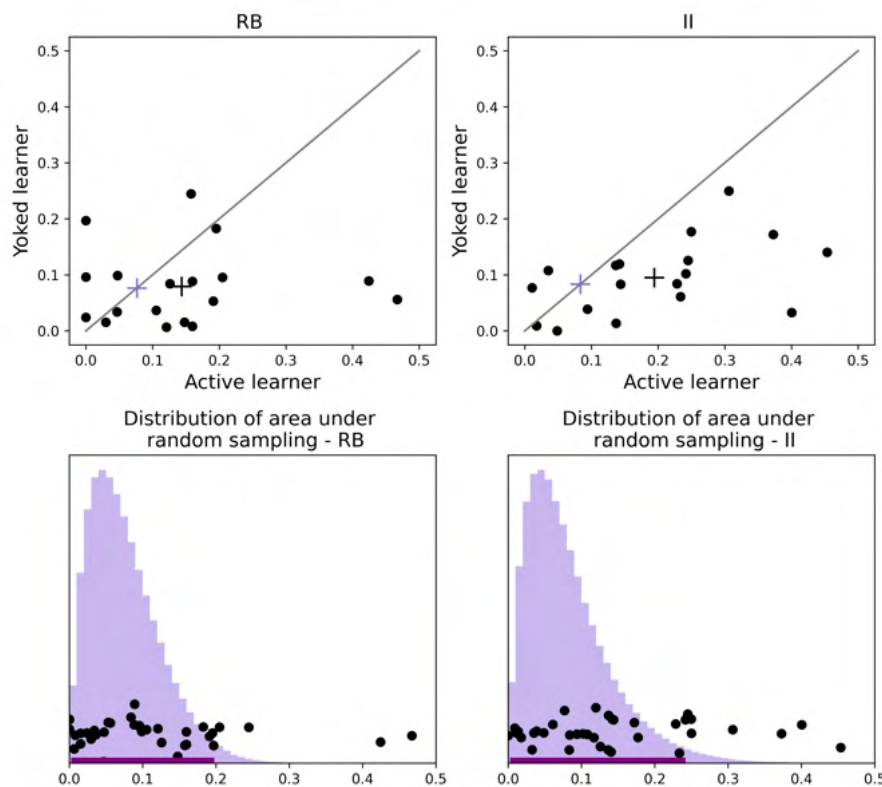


Figure 3.20: Upper panel: Area of polygons inscribed by the teaching examples, summed over the two categories (excluding overlap). Areas are scaled by the total stimulus space size. Each dot is a dyad. The black cross represents the observed sample means, and the purple cross represents the mean expected from random sampling. Given the structure of the II category, 6 examples can be sufficient to cover the entire stimulus space (for each category). However, the maximum area that can be covered in the RB condition with 6 examples and labels consistent with the true category is 50% of the total area. Lower panel: Distribution of area for random sampling of two triangles from the corresponding categories. Experimental data are overlaid in black (each dot is a participant; vertical jitter added for visualization).

Based on a pedagogical teaching intuition, we expected that there would be more variation in the feature irrelevant for categorization, than in the relevant one. At the individual level, the ratio of the variance of the orthogonal distances of examples to the boundary to variance to the orthogonal dimension was not significantly different than 1 in any of the groups [Figure 3.18](#), all $p > .5$. The mean value was consistent with the predicted direction.

The second measure considering a pedagogical learner as the reference, the smallest boundary distance, is presented in [Figure 3.21](#). This metric computed the summed orthogonal distance from the subjective boundary (inferred in the last test block) for the example closest

to the boundary from each category. Data from participants who only selected examples from one category was not included ($n=2$). The mean distance was smaller than expected by random sampling for both active and passive learners. However, there were no significant within-dyad differences in either condition, RB: $t(17) = -.72, p = .48, BF_{null} = 3.11$, II: $t(15) = .02, p = .98, BF_{null} = 3.91$.

The estimated 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: $r(16) = .05, p = .85, BF_{alt} = 1.99$.

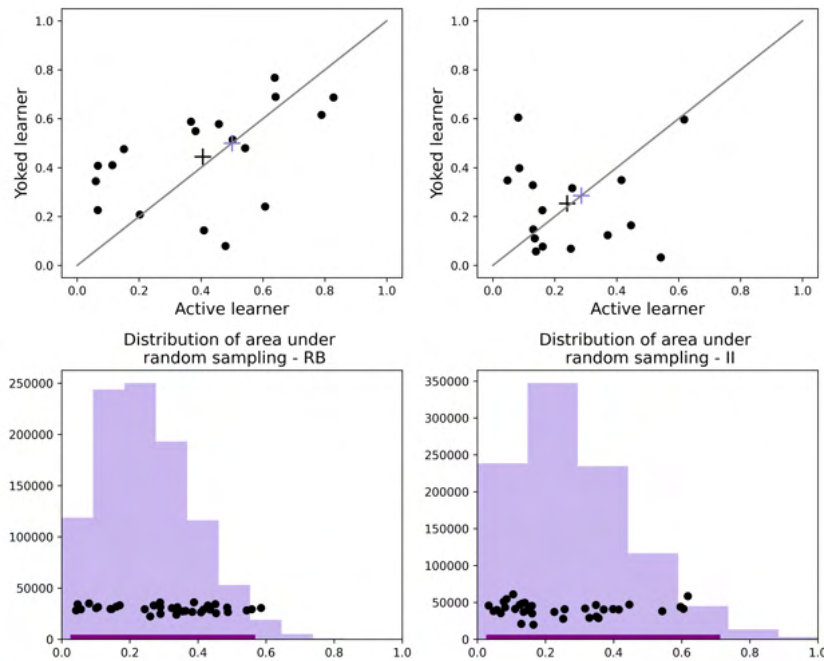


Figure 3.21: Upper panel: Distance to boundary of the two closest examples labelled assigned to different categories. Measurements were scaled to the unit square. Each dot is a dyad. Black crosses mark the observed sample means, and purple crosses represent the mean expected from random sampling. Lower panel: Distribution of metric under random sampling of three examples in each category.

Given that there significant within-dyad differences were found in categorization performance even in the last test block, it can be argued that the observed teaching differences were only an indirect effect of the active learners having a more accurate representation of the two categories. If categorization performance drives the results, we would expect that dyads that

have larger differences in terms of classification performance would also be the ones in which the largest differences are observed in teaching performance. The extent to which the active learner outperformed their yoked learner was correlated with each of the teaching metrics.

The difference in the number of consistent boundaries for dyad members correlated negatively (as predicted), but not significantly, with accuracy differences in the II condition, $r(14) = -.46, p = .07$. There was no correlation in the RB category, $r(14) = .16, p = .55$. Differences in the area inscribed by examples did not correlate significantly in either condition with accuracy differences, $RB : r(16) = .37, p = .12$; $II : r(16) = -.15, p = .54$, and followed the expected direction only in the RB condition.

The correlations in the II conditions were of medium magnitude, so underpowered according to post-hoc calculations. Therefore, a resampling test was performed to test whether the observed difference in the number of consistent hypotheses observed between II active and passive learners was entirely explainable by accuracy. 500 groups of 10 active learners and 10 passive learners were randomly sampled with replacement from the participant pool and were kept in the analysis only if their average accuracy differed by maximum 1%. For these sets of participants matched in accuracy, the mean difference in the number of consistent samples was computed. The mean difference observed across these samples was -.86, and the interval including 95% of the values was [-.89, -.81], excluding zero.

Inter-individual differences in active learning and teaching performance

To test whether better active learners were also better teachers, the average query distance of active learners was correlated with the subsequent teaching performance. There was a moderate significant correlation between query distance and the number of compatible boundaries after teaching in the II condition, $II : r(15) = .47; p = .05, BF_{alt} = 2.03$, as participants who made queries further from their subjective boundary, tended to be worse teachers. There was only a small, non-significant correlation ($RB : r(15) = -.16; p > .05$), for the query distance and teaching in the RB condition.

Figure 3.22 illustrates the relationship between accuracy, active learning performance, and the number of consistent boundaries after teaching. Accuracy correlated positively with teach-

ing performance, but did not reach the significance level ($RB : r(15) = -.39, p = .10$; $II : r(15) = -.44, p = .07$).

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.

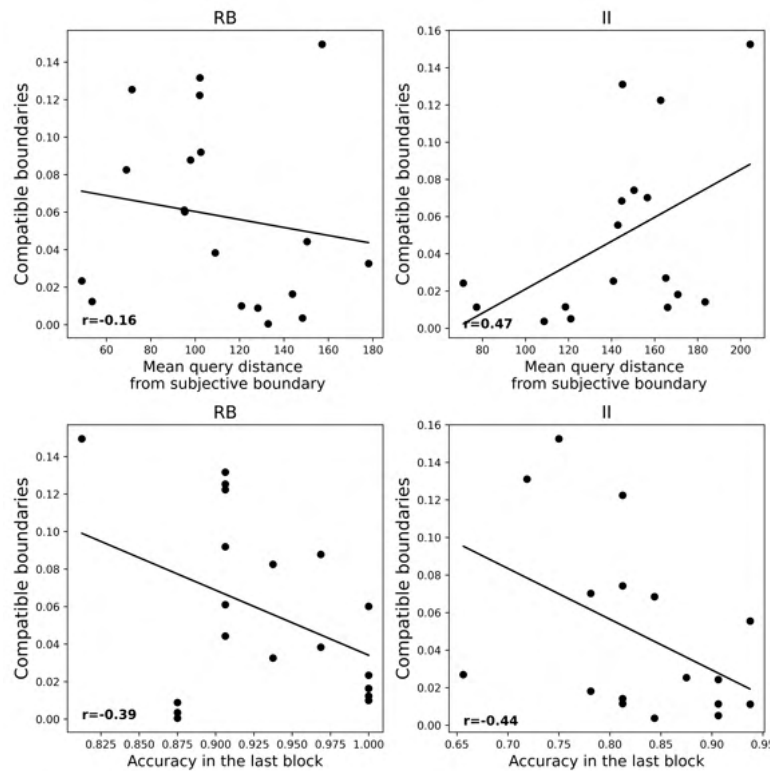


Figure 3.22: Upper panel: Relationship between active learning and teaching performance. Active learning is measured by the average distance of the queries to the subjective boundaries across blocks. Lower panel: Relationship between classification accuracy in the final block and teaching performance. Each dot is an active learner. Only teachers who provided linearly separable teaching sets were included. None of the correlations were significant.

Example order

Participants showed relatively agreement in choosing opposite corners of the stimulus space as the first two examples. Specifically, participants started by choosing two samples that maximally differed in both features. Colormaps of example choice frequencies, presented for every example order, are shown in [Figure 3.39](#).

In order to quantify potential order effects, the distance to the subjective boundary was computed for each of the six teaching samples as a function of the sample order. Since the

maximum distance that could be produced by a participant differed as a function of the boundary they tried to teach, the distances were rescaled for each participant to the [0,1] interval, such that 0 corresponded to the sample closest to the boundary and 1 to the sample farthest away from the boundary.

The participant pool was then split according to how they chose to order examples from different categories by calculating the number of category switches for every participant. The minimum is 0, which occurs when participants only chose examples from one category, and the maximum is 5, when participants alternated each subsequent category assignment.

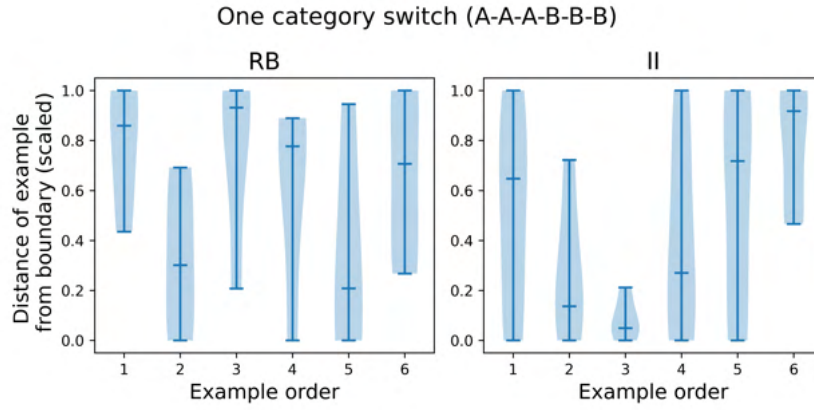
Five category switches should be easily amenable to a curriculum learning strategy, first presenting examples on both edges of the space that are unambiguous and easy to categorize and then narrowing into the region surrounding the boundary. [Figure 3.23](#) suggests that curriculum teaching is compatible with the distribution of examples in II structure, but there is no (visually) discernible pattern in condition RB.

Further, participants who made only one category switches seemed to follow a linear strategy in the II condition ([Figure 3.23](#)), traversing the stimulus space from one edge to the other. Again, the pattern for the RB category shows less consistency.

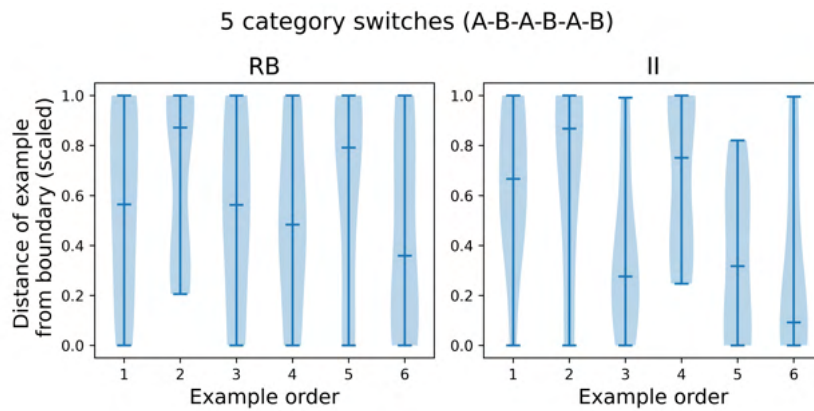
Given the small subsample size, we did not perform any statistical tests and the observations are purely descriptive.

Qualitative data

We do not present an analysis of the open ended answers of the participants about the teaching phase since very few of them wrote down meaningful explanations for their choices that could be coded for analysis. Anecdotally, the majority of explanations for example choices were consistent with an attempt to provide exemplars covering the entire stimulus space (e.g. "I wanted to select all of the possible kinds of antennas that receive each channel"). Some participants in the RB condition explicitly pointed out the perceptual feature that the categorisation was about, while others in the II condition said they would have mentioned to the learner that both features matter. Some participants in the II condition mentioned avoiding offering as examples antennas that they were uncertain about themselves to avoid confusing the learner.



(a) Subset of participants who consecutively chose three examples of the same category followed by three examples of the opposite category. Very few participants chose this strategy: $n=4$ for the RB structure and $n=5$ for the II structure.



(b) Subset of participants who alternated the category of each successive example. This was the most common strategy (but still accounting for only 27.50% of the full sample), $n=12$ RB structure and $n=10$ II structure participants fell in this category.

Figure 3.23: Violin plots showing the distance to the subjective boundary as a function of example order. Distances have been scaled to the 0-1 range for every participant. Vertical lines mark medians. Active and yoked learners are pooled together within a category structure.

3.4.4 Conclusion

We closely replicated the results of Markant and Gureckis (2014): all active learners were able to learn the RB category structure, whereas many active learners failed to learn the II structure. The accuracy of II active learners was higher than in Markant and Gureckis (2014), as about 65% of participants scored higher than expected based on the use of only one classification feature in the final training block. However, on average, the queries made by II category learners failed to converge around the true boundary. In line with previous results, yoked passive learn-

ers had lower classification accuracy in both conditions and they did not benefit from being yoked to better active learners.

In order to assess teaching performance we computed two metrics for teaching which should rate the usefulness of teaching examples to a naïve learner, who does not know they are being taught and who has little perceptual noise. The first quantifies the amount of uncertainty that this learner will be left with about the boundary separating categories, after excluding boundaries incompatible with the teaching set presented. The second computes the area of the stimulus space inscribed by examples. Learners exposed to teaching examples that span a larger area will be familiar with the more exemplars from the categories.

Further, based on simulations from the rational pedagogy model, we created two more metrics capturing potential intuitions of learners who know they are in a pedagogical context. First, we calculated the variability in examples across the dimension relevant for categorization and contrasted it to variability in the dimension irrelevant for categorization. This metric is also supported empirically by findings that increased variability in the irrelevant dimension improves learning outcomes in RB categories (Rosedahl & Ashby, 2021). Very interestingly, the irrelevant variability dissociates RB and II categories in terms of learning outcomes as Rosedahl and Ashby (2021) recently showed that it impairs the learning of II categories. Second, we calculated, for each category, the orthogonal distances from the closest examples to the subjective boundary. Here the intuition was that learners faced with teaching sets that are linearly separable with respect to several boundaries, will assume that the boundary intended is the one that minimally separates the two categories.

The results supported the main hypothesis of differences in teaching performance as a function of having active vs. passive learning experience. Active learners provided examples for teaching that more narrowly constrained the set of possible hypotheses in the II condition, but not in the RB condition. This is in line with the expectation that differences between the active and yoked teachers would be higher for the II structure which was generally more difficult. Similarly, differences active learners inscribed a wider area of the stimulus space with their examples than the yoked controls. We found no significant difference for RB learners.

This is likely due to the fact that yoked RB learners likely randomly sampled the values of the categorization-irrelevant feature of the stimuli from the category range.

Moving on to the pedagogical teaching measures, we found no difference in terms of the use of irrelevant variability in either condition. In fact, variability across the relevant and irrelevant features was comparable.

There was no significant within-dyad difference in how close participants placed examples to their subjective boundaries. One potential concern with this metric is that it relies on having a good estimate for the subjective boundary used by teachers. It is important to show that the subjective boundary of participants modulates the chosen teaching examples, however, the downside is that any bias in the estimation of the boundary will be translated into this teaching metric. However, we found no significant differences on this distance metric in the RB condition where the determinism of the boundary was high and boundaries were stable over the last blocks and, as a consequence, we can be fairly certain that the boundary estimated was unbiased.

The secondary aim of the experiment was to control whether the benefits of active learning on subsequent teaching were fully mediated by having higher accuracy in the categorization task before teaching. There was a medium strength (but statistically non-significant) positive correlation between differences in categorization accuracy within a dyad and differences in teaching performance. However, across resampled groups of active and yoked learners matched for accuracy, we still found significant differences in teaching performance. While this is not a within-dyad analysis, it suggests that the mode of learning influenced also learning directly and not just indirectly through categorization accuracy. In line with this, we found that the partial correlation between query distance during active learning and teaching performance (quantified through the number of compatible boundaries), given accuracy in categorization, was again of medium magnitude and in the predicted direction (though non-significant) for the II condition.

Third, we found a medium (statistically non-significant) correlation in the predicted direction for the determinism of the boundary during categorization and the distance of examples to the boundary for active learners. Given the lack of statistical significance, we cannot draw any

conclusions currently, but given the effect size, further investigation is warranted. It is certainly encouraging that even in such a complex perceptual task, teachers were able to take their own uncertainty into account. In general, 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. It is possible that alongside their own uncertainty, teachers additionally took into account the potential perceptual uncertainty of the learner, thus further increasing the need for distinctive teaching examples.

It would be particularly interesting to test, in the context of an interactive teacher-learner experiment using a similar paradigm, whether teachers can additionally modulate example set selection based on the inferred uncertainty of their learner, or whether they are only influenced by their own uncertainty.

While teachers converged on choosing corners of the stimulus space as the first two examples, there was no strong agreement on continued curriculum teaching, that is, starting with examples that are easier to categorize and gradually providing examples closer to the boundary, whose category membership is more uncertain. This is surprising especially for active learners who themselves started by querying the limits of the stimulus space.

A limitation of the current experiment is that the utility of the examples provided by teachers was quantified by metrics that should in principle guarantee better performance for learners, but that certainly do not map one to one to the practical benefits derived by actual learners.

One of the very few experiments on teaching can shed some light on this issue. Interestingly, in reverse logic to the current experiment, Avrahami et al. (1997) used teaching by example giving as a paradigm to uncover people's intuitions about category learning under the assumption that teachers were giving the best possible examples for learning. Participants in their study learned categories based on a small set of the (Nosofsky, 1989) stimuli (20/36 stimuli) drawn together on a piece of paper. They were only allowed to start teaching once they had perfectly mastered the category to be taught. Teachers could select examples only for one category from the same preset list, but were allowed to give as many examples as they wished. Avrahami et al. (1997) found high consistency between the examples offered by participants,

but no effects of the way in which the participants had learned the category (by verbal description, by having them marked, but asking about the category) or the mode of teaching (to an imagined student or to a real passive student). The lack of differences as a function of the mode of learning are not surprising in this context given that the ‘selected exemplars’ learning mode entailed participants asking about the membership of all items presented, rather than actually engaging in self-directed learning. Avrahami et al. (1997) computed the most common teaching sequence (choosing the most common example at each time step, conditional on the previously chosen example) and presented this sequence to new participants ($n=10/\text{condition}$) who acted as learners and then were tested on 8 new stimuli. It is unclear whether these participants were aware of the way in which the data were produced or not. In addition to the observed teaching sequence, Avrahami et al. (1997) also showed participants examples that came from two strategies: “separate dimension strategy” (examples should have the same value on the classification dimension and span maximally on the irrelevant dimension) and “borderline strategy” (chose examples closest to the boundary on either side and vary location along boundary in irrelevant direction). The borderline strategy corresponds to the pedagogical strategy we also outlined. Learners who were shown the teacher’s examples performed somewhat better on average than the participants shown examples conforming to the two strategies.

Given that these benefits exist, there are multiple possible explanations. First, it is possible that, as a function of the pedagogical assumptions of learners, some of the samples deemed less useful by the formal analysis are actually sufficient to convey the two categories. On the other hand, we could not show that teachers successfully applied the two intuitive pedagogical strategies that took into account the recursive aspect of teaching. Further, it is not immediately obvious that learners would be able to utilize information about the way in which examples were sampled since the informed yoked learners (who knew they were yoked to an active learner) in Markant and Gureckis (2014) did not perform any differently than the naïve yoked learners. Either way, to determine how learners make inferences based on pedagogically sampled data in this task, an experiment would need to be conducted in which the examples generated by teachers are shown to learners who are told that samples come from a random

(strong sampling) generation process or a teacher.

Second, potential learners may learn better than expected from the teaching examples produced by teachers due to a shared bias in category learning. Primarily, while we considered only whether a boundary is compatible or not with the dataset, it is likely that participants do not consider a wide range of possible boundaries and have priors about the types of category structures they expect. This is indeed what can be concluded based on the queries selected by active learners. Based on this, one can predict higher rates of success for learners on RB category structures, whom will have hypothesis spaces more closely aligned to those of their teachers.

In summary, we found evidence that active learning improves teaching in a complex task in which the category structure was not immediately apparent to participants (II condition), but failed to find any statistically significant differences in a task that was easier and in which the category structure was easily verbalizable (RB condition). We cautiously argue that the contribution of active learning, while mediated to some extent by final accuracy performance, had an independent contribution to the observed effect.

3.5 General Discussion

Across two experiments which shared a conceptual task, but differ greatly both in difficulty level as well as methods, we show that teachers benefit from active learning experience.

In Experiment 1, as well as in the RB condition of Experiment 2, we did not find a significant effect of active learning experience on teaching performance within dyads of active and yoked passive learners. In both situations, the effect observed was in the predicted direction, but was very small. On the other hand, we did find a significant effect of learning mode in condition II of Experiment 2. This difference maps to an important distinction in the experiments, namely that in both Experiments 1 and condition RB of Experiment 2, the categorization boundary that participants taught was verbalizable, and strategies for teaching could be explicitly transferred from the learning task. This resonates with the idea that active learning

is particularly useful, beyond passive learning, when the structure of the generating model for the stimuli is not immediately available.

Related to this, in a recent paper, Rosedahl et al. (2021) found that explicit written and verbal instruction did not improve performance in the II category learning task, but did so for RB category structures. It is surprising that explicit knowledge about the structure of the hypothesis space of II categories, specifically the fact that both stimulus features are required for categorization, did not help participants. Rosedahl et al. (2021) fit their results within the COVIS (COmpetition between Verbal and Implicit Systems, Ashby and Maddox (2005)) model for categorization. Since RB category learning according to COVIS relies on an explicit rule-discovery process, it is easy to see how explicit instruction about the relevant rule is useful. On the other hand, if II category learning relies on procedural learning that is not conscious, knowledge about the rule is not going to be as applicable during learning. In light of this result, it would be particularly interesting to check whether instruction by curated examples would improve the performance of learners in the II condition. If that is the case, it would be a good argument for the relevance of teaching-by-induction in ecological settings.

Indeed, the participants in Experiment 2 found it impossible to verbalize their teaching strategies, despite the fact that some of them near followed near optimal teaching strategies in the perceptual space. It would have been interesting to ask these participants to perform the exact same task in a conceptual space to check whether they would have the same intuitions for solving the teaching task.

Shafto et al. (2014) found that, if learners were first teachers, their behavior was better accounted for by the rational pedagogy model. To explain this, they speculate about a process-level account according to which the large recursive inference demands are side-stepped by using pre-computed values from teaching in the learning.

To our knowledge, Experiment 2 is the only lab based study of teaching-by-example where learning is extended in time and, therefore, teachers have imperfect knowledge of the teaching material. This departs from the teaching games framework used previously by Shafto et al. (2014) in which the teacher can easily imagine their reasoning process had they been a taught

learner, making the solution to the recursive problem easier to find. This should explain why we found teaching behavior consistent with generic principles of good teaching (e.g. providing unambiguous examples), but not with specific rational-pedagogical predictions. However, it is possible that with more training, and less noisy boundaries, the predictions of rational pedagogy will pan out.

The next questions in this line of research should target the flexibility of teaching strategies. Particularly, the correlation between the teacher's subjective uncertainty and sampling behavior should be confirmed. Further, in an interactive design, where the teacher is allowed to observe the learner, it would be possible to ask the extent to which the teacher flexibly adapts to the uncertainty of the learners and their current state of knowledge.

3.6 Supplementary Information

3.6.1 Pilot: Priors about boundary locations

In our computation of expected information gain, which determined the optimal query choice during active learning, we assumed boundaries are equally likely a priori. However, it is possible that, even for intentionally arbitrarily designed cover stories, participants might have non-uniform priors over the locations of the boundary between the two categories is located.

There are two potential solutions to this problem. The first is to try to estimate the participants' prior over the potential boundary locations. However, this can be challenging. First, it is possible that explicitly asking a participant questions about the probability associated with each boundary location is going to interfere with performance in the following learning and teaching tasks. On the other hand, using priors inferred by aggregation from one group of participants and extrapolating them to the participants completing the experiment is also unsatisfactory as there might be significant inter-individual differences.

We opted for the second solution which is to select stimuli and classification dimensions such that participants are likely to have a uniform prior over the possible boundary locations. A small pilot experiment was used to test whether the stimuli/categorization dimensions considered for inclusion in the experiment indeed elicited uniform priors over the possible boundary locations.

| Dimension | Boundary (less or more than...) |
|--|--|
| speed of animals relative to their body size | speed of the average human |
| loudness of musical instruments | 85dB (threshold for hearing damage) |
| date of domestication of plants | orange tree domestication |
| carbon footprint of food items | 15minute car ride |
| average sleep time per day | average human sleep time |
| vitamin B content | daily recommended dose |
| color | red/blue |
| price | cheap/expensive |
| shape morph | square/circle |

Table 3.1: Dimensions and boundaries used in pilot for Experiment 1

Methods

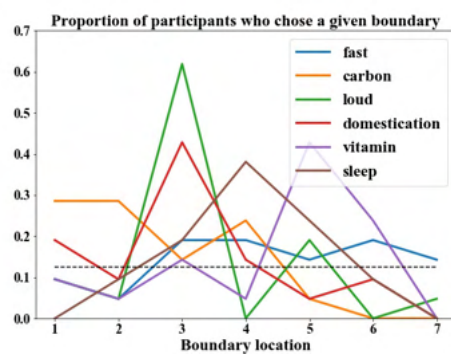
A list of nine classification dimensions and associated sets of 8 images each were compiled. The image sets were selected, without any overlap, from MultiPic databank of standardized color drawings of concrete concepts (Duñabeitia et al., 2018). Classification dimensions were associated with a seemingly objective boundary (see Table 3.6.1).

Participants were asked to choose where they thought the boundary between the categories lies based solely on their prior knowledge. At the end, participants also were asked to verbally explain their choices.

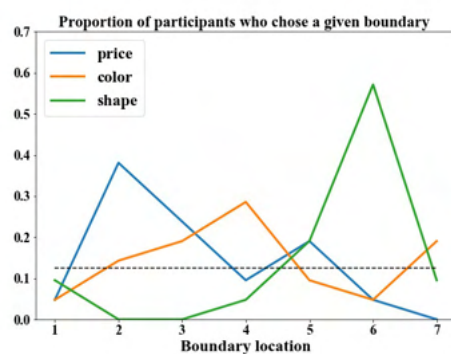
Twenty-one participants with native or good English proficiency were recruited through the CEU Research Participation System and were rewarded for performance with vouchers.

Results

Figure 3.24 illustrates the participants' aggregate intuitions about the boundary locations, and how they differ from the uniform baseline. A chi-square goodness of fit test for the uniform distribution was used to determine dimensions that needed to be discarded or modified. Some dimensions were excluded outright (price, color, and shape), and for other dimensions, stimulus substitutions were made (loudness, vitamin, domestication) based on the open-ended reports of the participants.



(a) Dimensions selected for inclusion in the experiment.



(b) Dimensions discarded in future experiment.

Figure 3.24: Inferred participants priors over the possible boundary locations.

3.6.2 Experiment 1: Additional analyses and figures

Optimality in active learning

Performance in active learning was contrasted to optimal behavior, choosing a query that maximizes information gain. Participants appeared to be near-optimal in the selection of the first query. Conditional on the first query, little more than a third of participants selected the maximally informative query. A large proportion of choices were close to the optimal solution (Figure 3.25), but another third made a completely uninformative query.

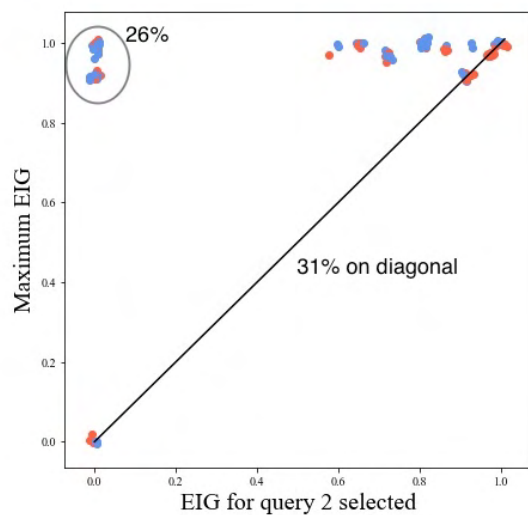


Figure 3.25: Expected information gain from the second active learning query against the maximal information gain given the first query. Each dot is a query selection. Colors denote conditions.

Contrasting teaching performance to chance

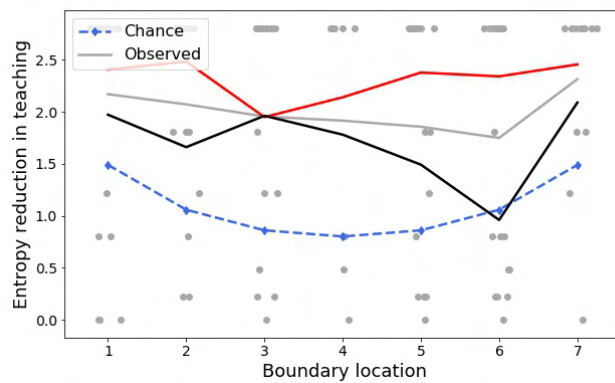


Figure 3.26: Chance level for entropy reduction during teaching as a function of boundary location (blue line) against observed performance (gray line) and performance broken down by group, with prior active learning in red and without prior active learning in black.

Decisions about boundary location

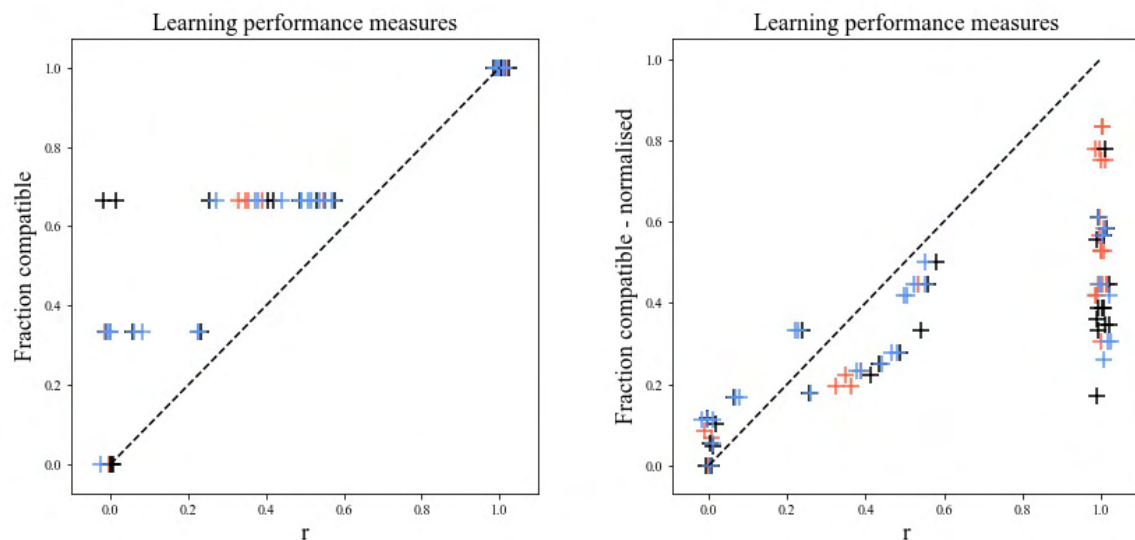


Figure 3.27: Estimated r against proportion of decisions compatible with the labelled data (left) and against normalized proportion of decisions compatible with the labelled data (right). Each cross represents a participant and conditions are denoted by colors.

We considered three possible ways of quantifying the performance of decisions about the boundary location: the proportion of selections (out of 3) that were compatible with the labelled data observed, the same proportion normalized by the number of compatible boundaries, and the estimated r parameter described in the main body. Figure 3.27 shows how these measures relate to each other. The proportion of compatible boundaries is too liberal and the normalized version too conservative, with the estimated r lying inbetween.

There were significant differences within dyads in decisions about the boundary locations according to all measures (see Figure 3.28). However, the performance on boundary selection was not correlated with teaching performance (see Figure 3.30), and any differences observed between groups cannot explain the teaching differences between groups.

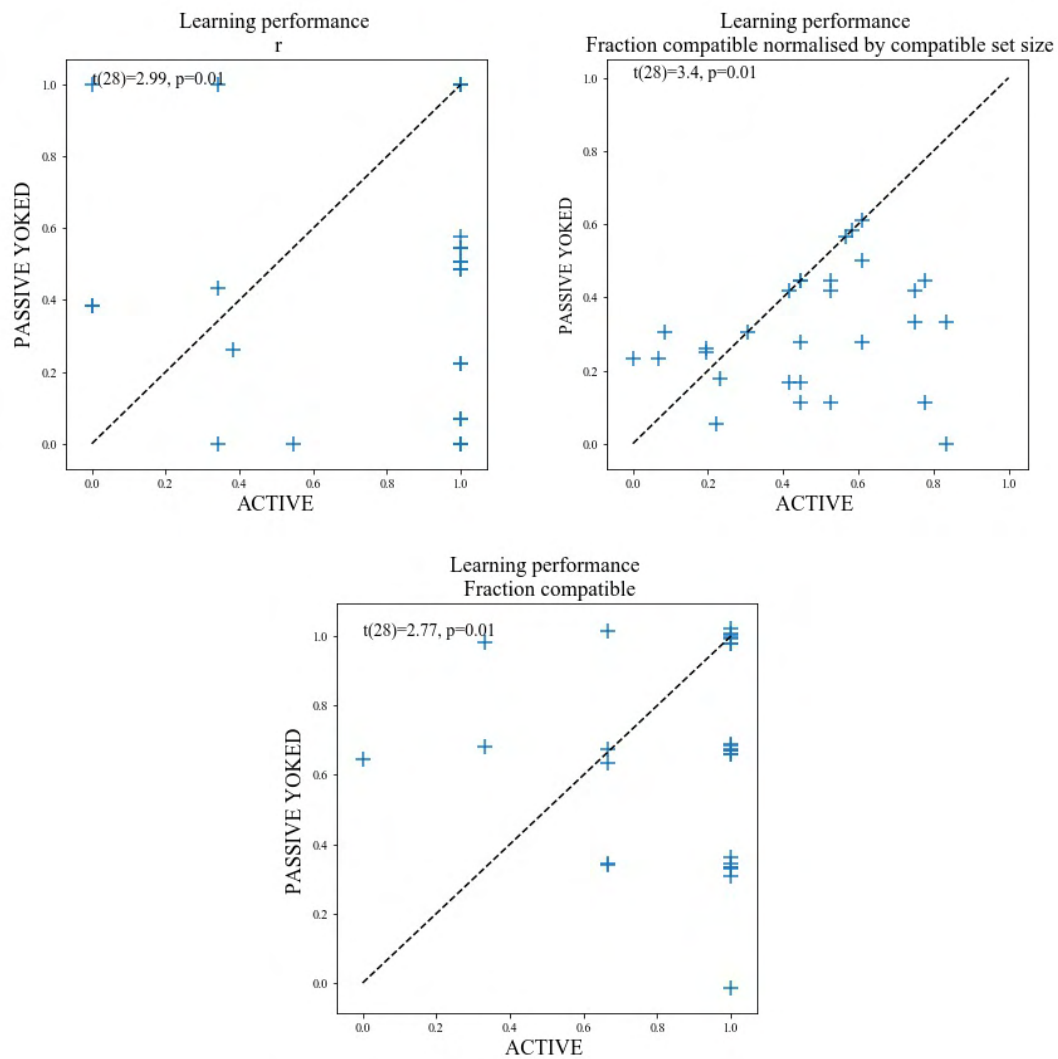


Figure 3.28: Within dyad performance differences. Each cross represents a participant.

Figure 3.29: Estimated r against proportion of decisions compatible with the labelled data (left) and against normalized proportion of decisions compatible with the labelled data (right). Each cross represents a participant and conditions are denoted by colors.

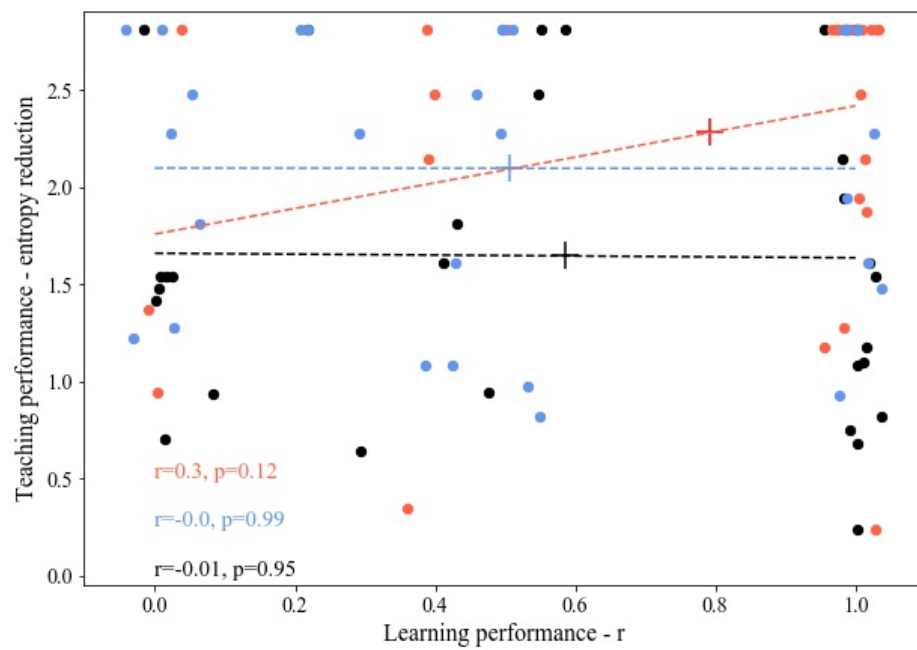


Figure 3.30: Teaching information gain as a function of the r parameter. The fitted OLS regression line for each condition.

3.6.3 Experiment 2: Additional analyses and figures

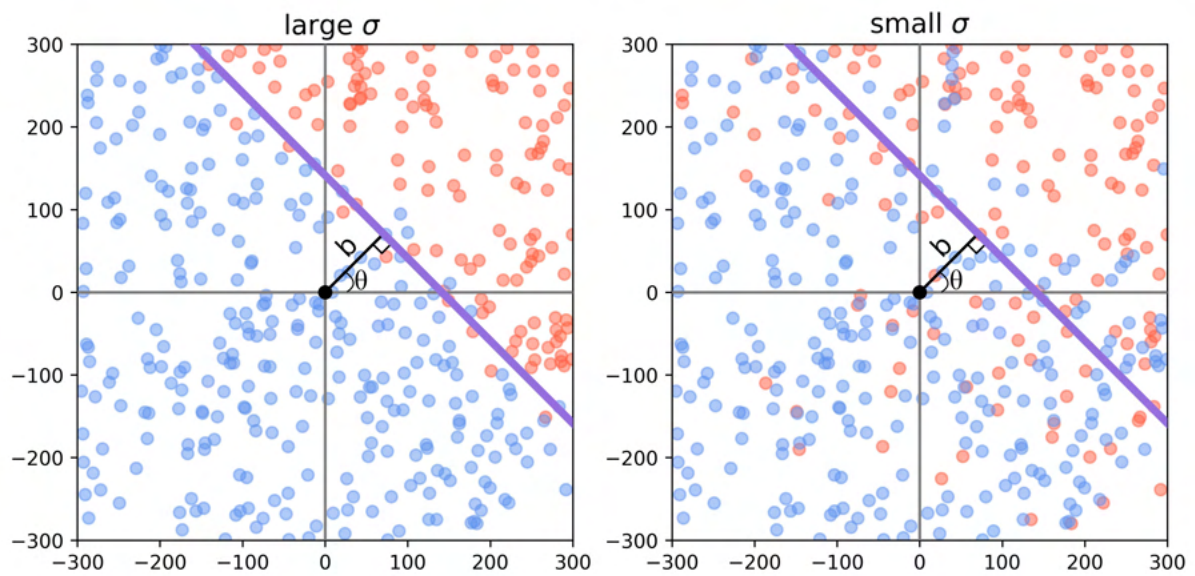


Figure 3.31: Visual illustration of the decision-bound model used to fit the participants' subjective boundaries (shown in purple) from their category choices.

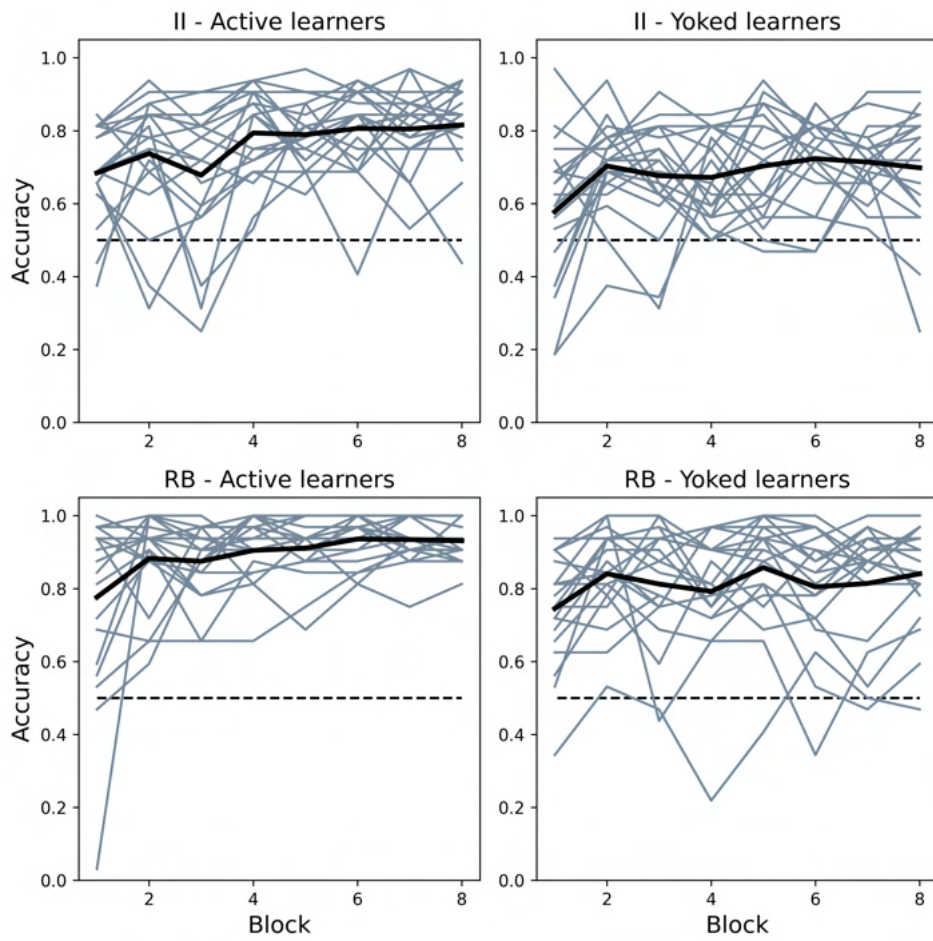
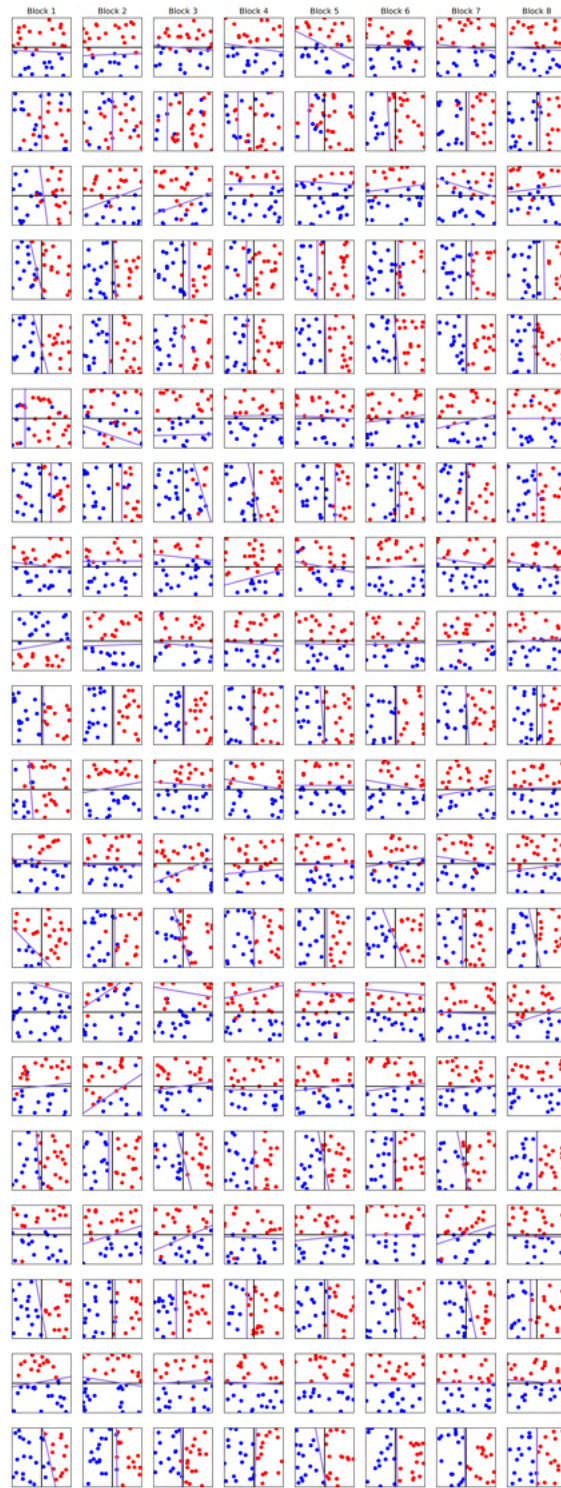
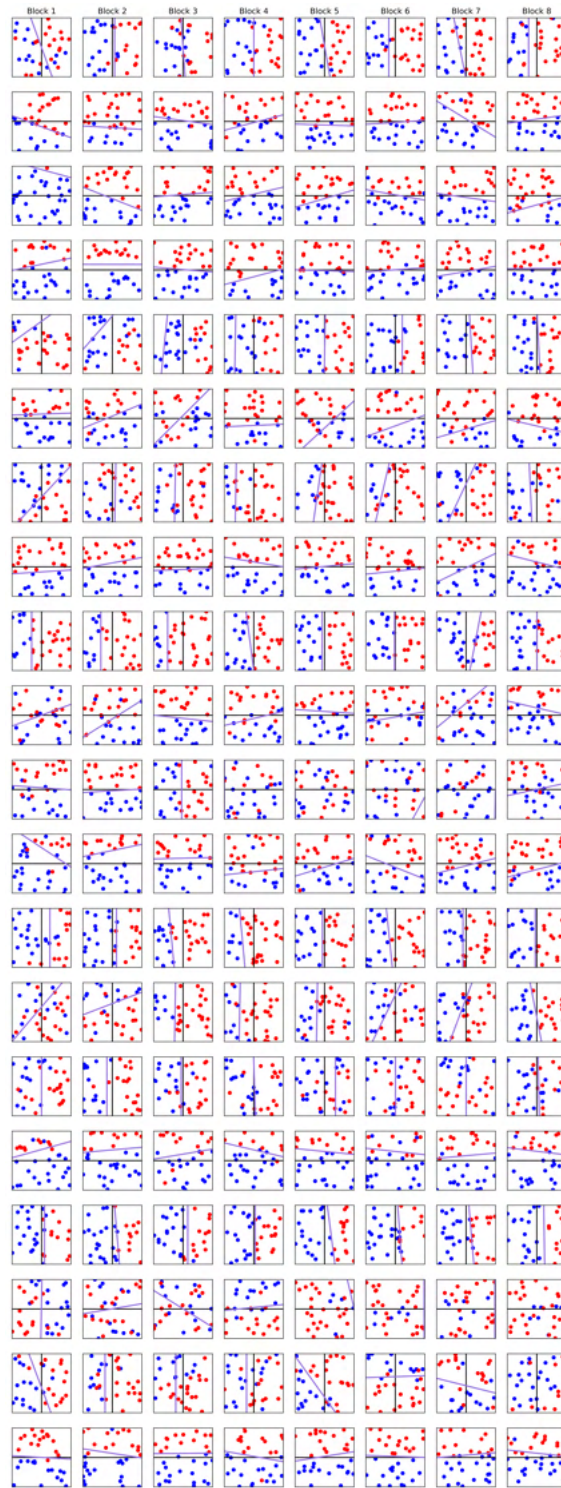


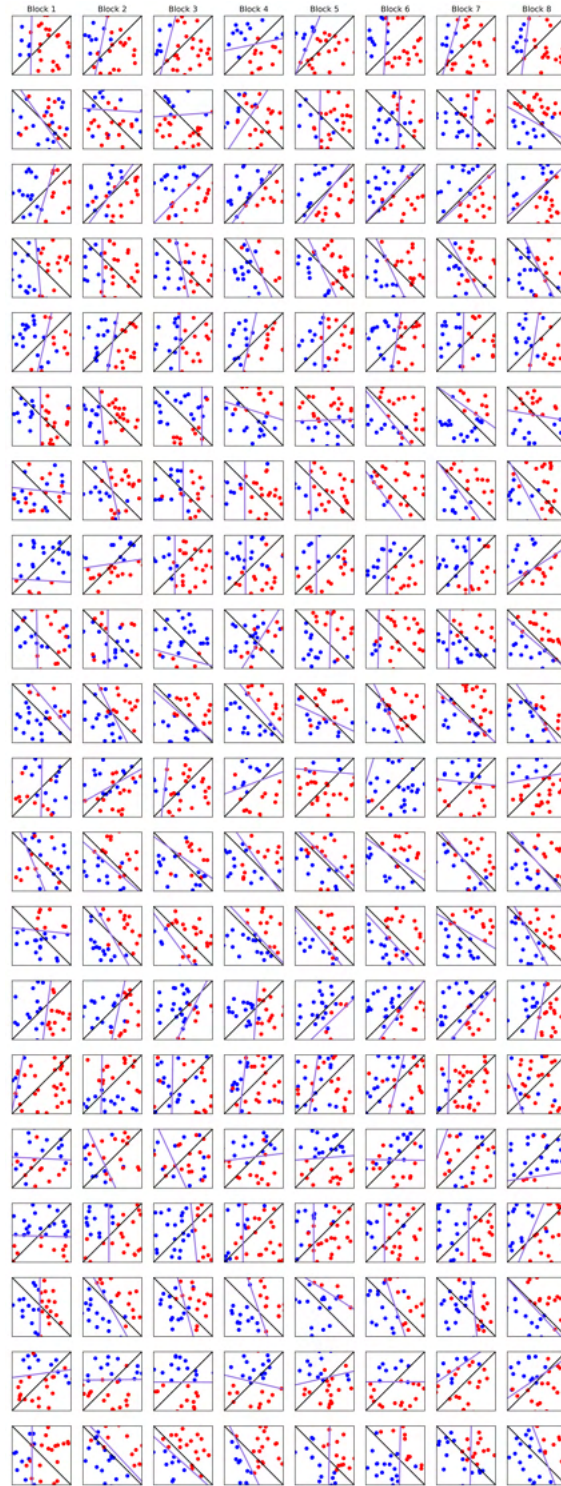
Figure 3.32: Categorization accuracy across participants and blocks.



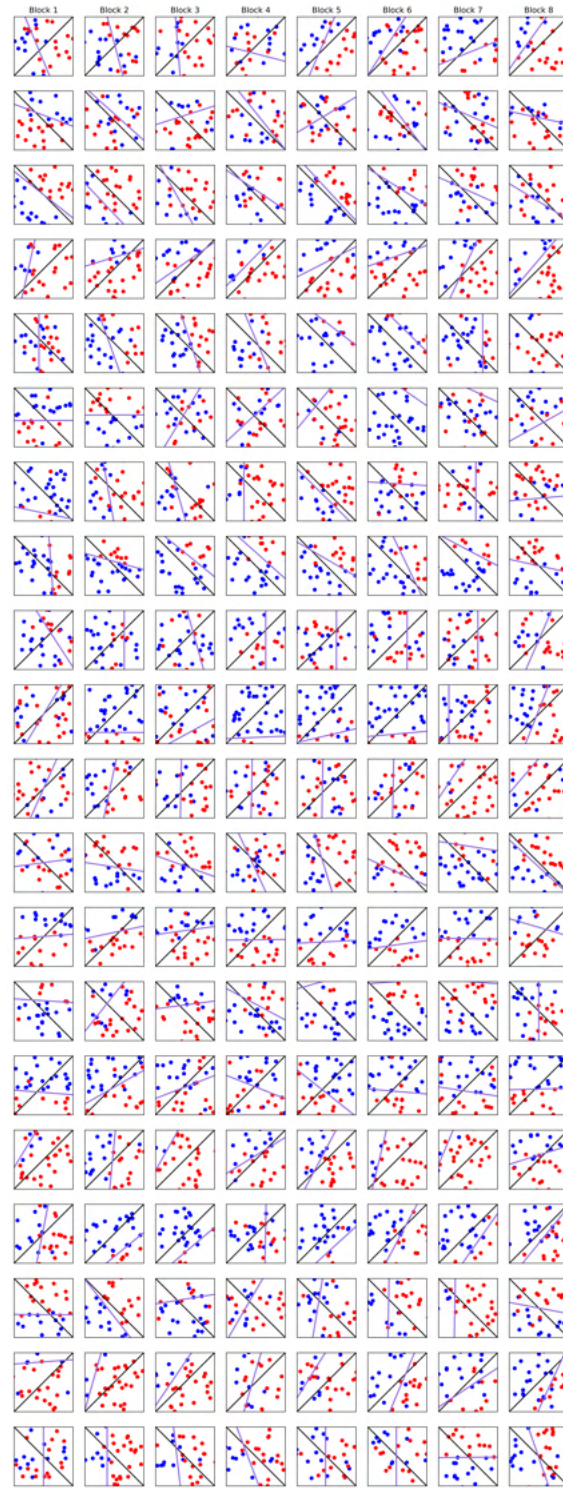
(a) RB - Active learners



(b) RB - Yoked learners

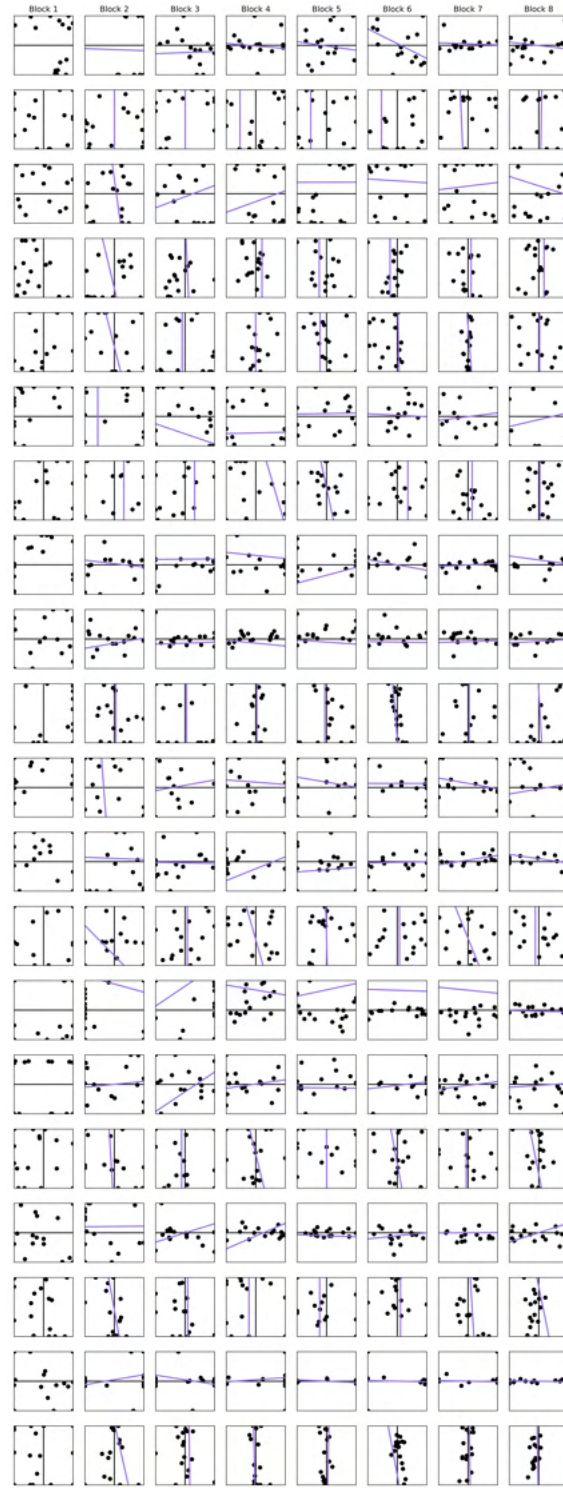


(c) II - Active learners

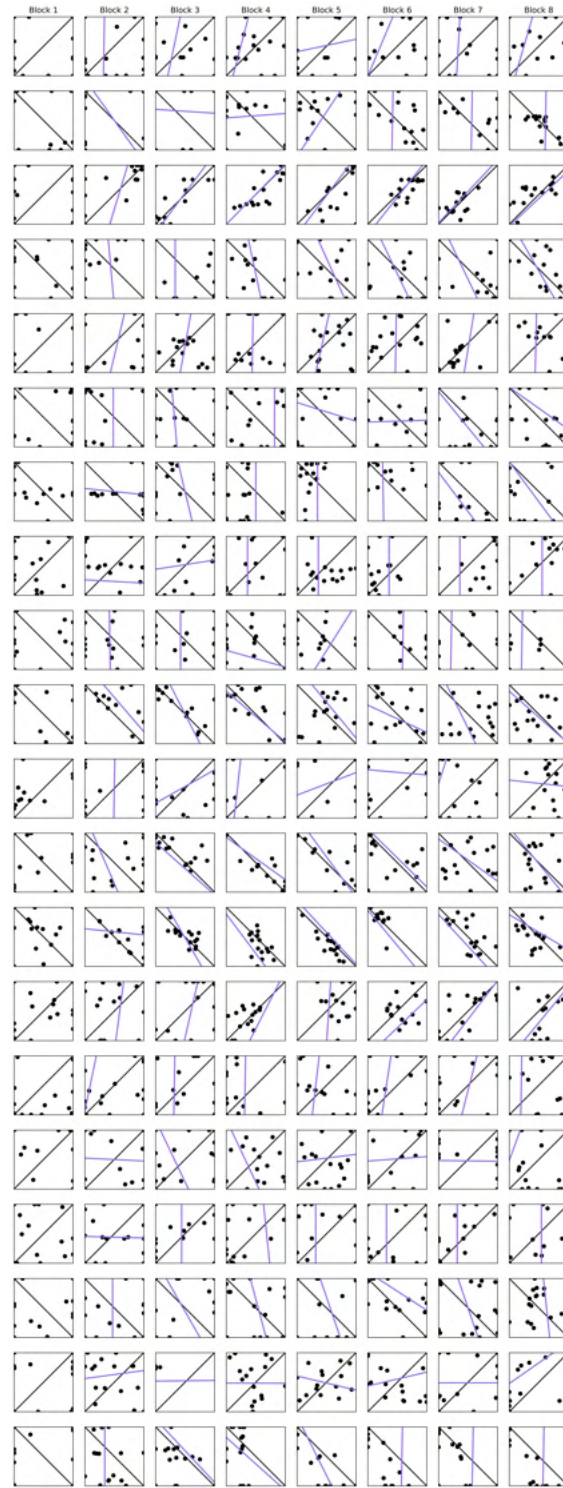


(d) II - Yoked learners

Figure 3.33: Categorization. Each row is a participant.



(a) RB - Active learners



(b) II - Active learners

Figure 3.34: Active learning. Each row is a participant.

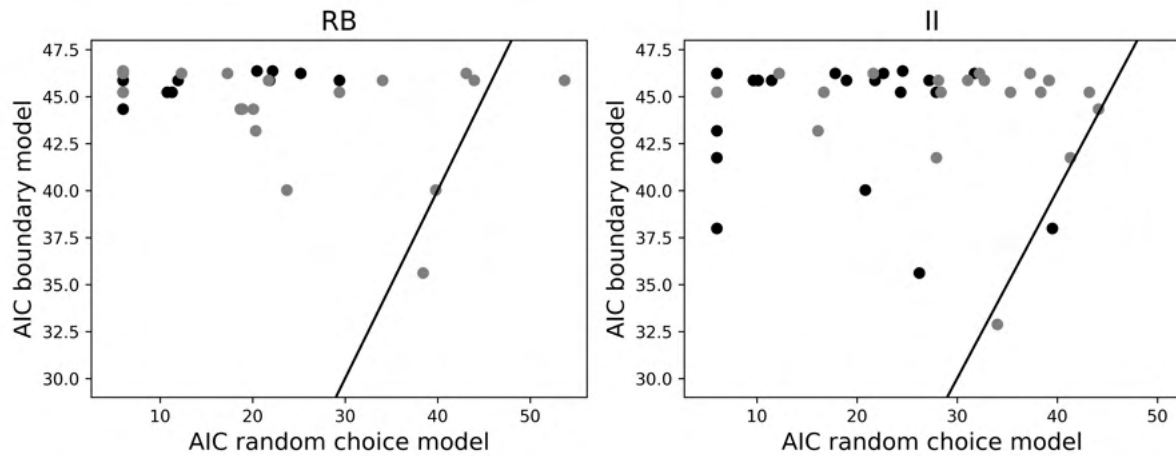


Figure 3.35: AIC values for a strong sampling response model (fitting just the probability of responding with a certain category) and the decision-boundary model used in the analysis. Each dot is a participant. Black dots represent active learners and grey dots represent yoked learners.

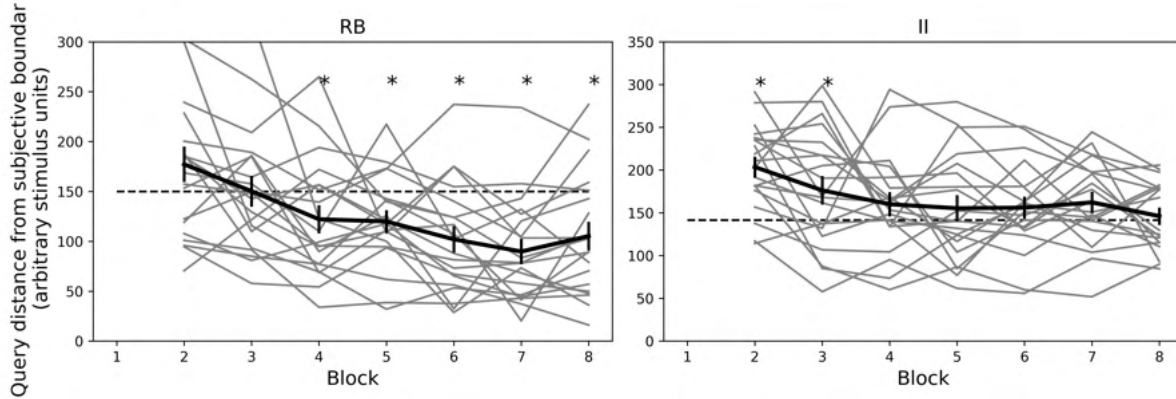


Figure 3.36: Average query distance to (individual) subjective boundaries. Bars represent standard errors of the mean. The dashed line is the expected query distance under strong sampling for each corresponding category structure. Asterisks are displayed for blocks where the average query distance differs from the dashed line significantly in a one-sample two-tailed t-test.

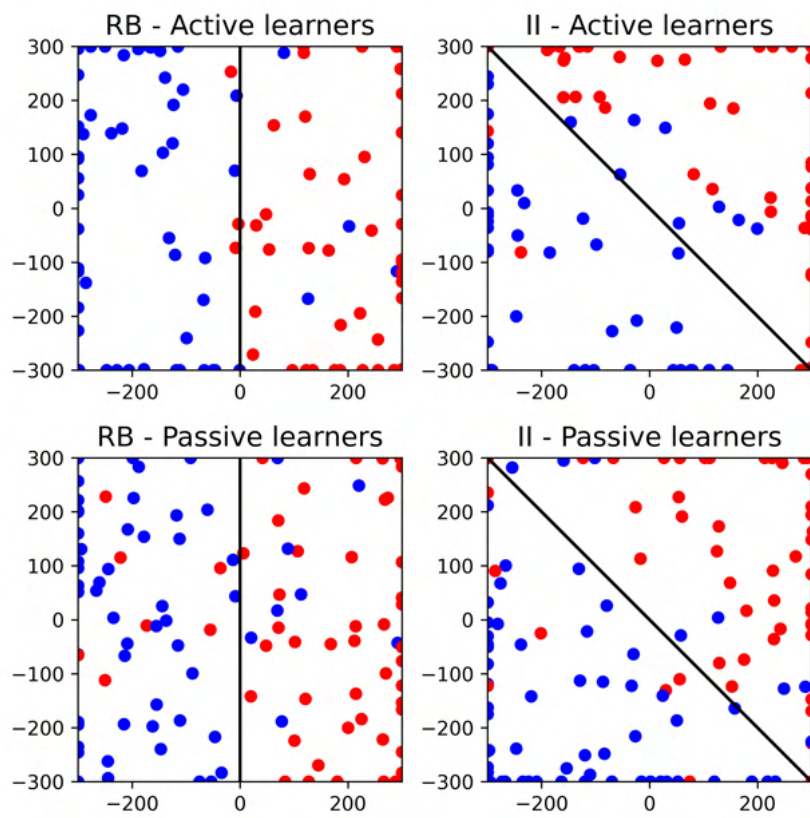


Figure 3.37: All examples chosen by participants. Colors denote the categories chosen by the participants before designing the example.

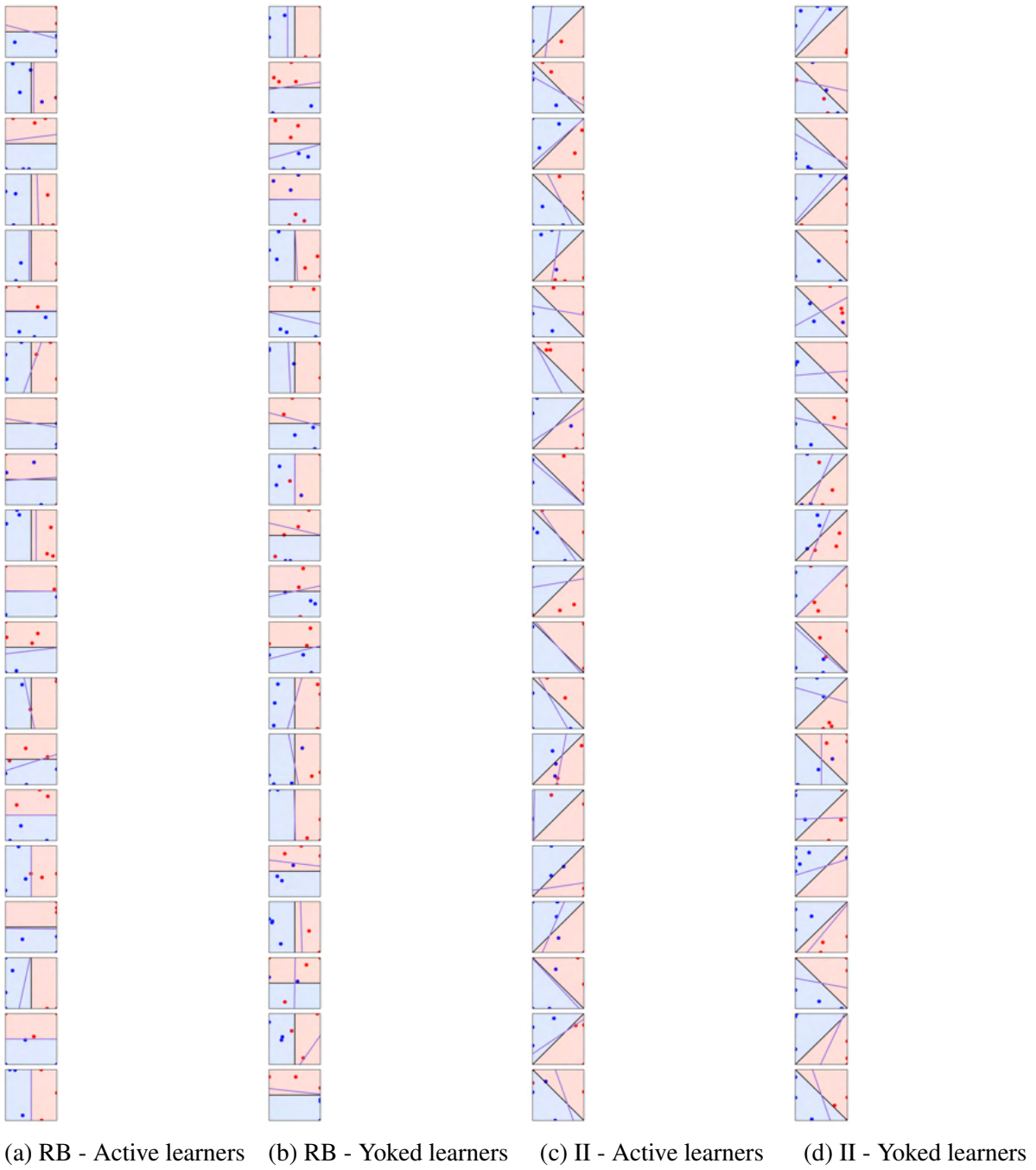


Figure 3.38: Teaching. Each row is a participant. Area colors mark the correct category. Dots are stimuli offered as examples. The black lines are the true boundaries and purple lines are the subjective boundaries inferred from the last test block.

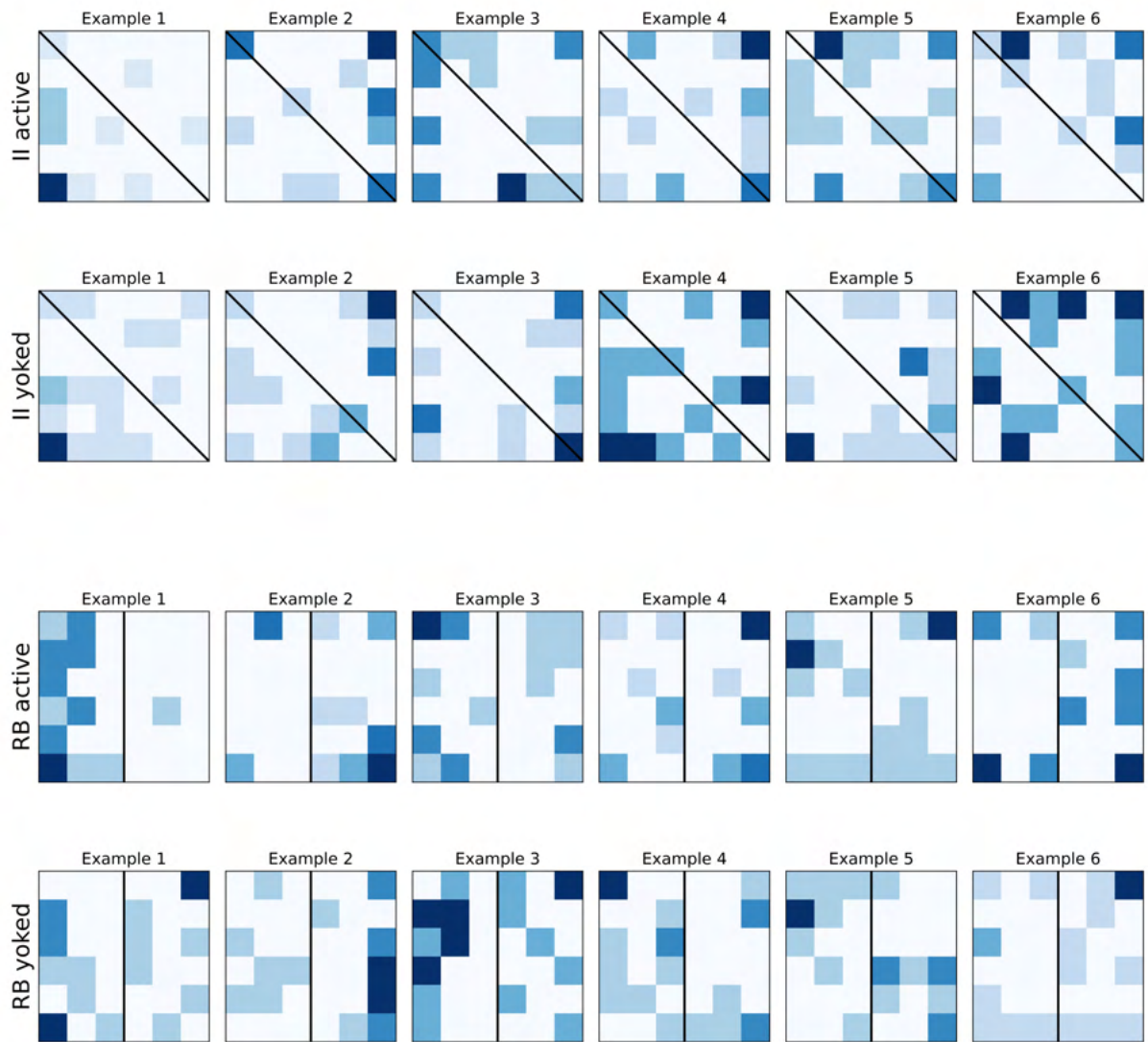


Figure 3.39: Frequency of teaching examples across the stimulus space (pooled across all participants). Color intensity corresponds to the number of examples in each bin. Each bin contains about 3% of samples under uniform sampling.

Comparing distribution of teaching examples to the uniform

To determine whether the distribution of examples was significantly different from random uniform sampling of the stimulus space, the one-sample Kolmogorov-Smirnov (KS) distance (the largest distance between the cumulative distributions of the observed data and a reference distribution) was computed between the observed distribution and the uniform. Distances were highly significant for both active (RB: $D = .92$, $p_i .001$; II: $D = .81$, $p_i .001$) and yoked learners (RB: $D = .97$, $p_i .001$; II: $.92$, $p_i .001$)¹².

To compare the relative extent to which uniform sampling can approximate the observed pattern of results in the two category structures, we also computed the Kullback-Leibler divergence between the samples. The KL divergence measures the amount of entropy remaining in a probability distribution when approximated by another one. While the measure is not symmetric, it has the useful property that it only takes the value of zero if the two distributions are equivalent. The KL divergence was larger for active learners (RB: $KL = .40$; II: $KL = .57$ nats) than for passive learners (RB: $KL = .25$; II: $KL = .35$ nats) in both conditions. The distribution of the test statistic under the null was computed by sampling¹³. The interval containing 95% of values was $[.09, .24]$, and excluded all observed KL values.

The difference between conditions may be explained by the fact that participants in the RB condition were (intentionally or unintentionally) randomizing values sampled for the irrelevant stimulus feature. To test for this, two-sample KS distances were computed between the observed distribution and a distribution that preserves the marginal distribution across the relevant feature, but uniformly samples the value of the irrelevant feature. The KS test suggested that the irrelevant feature was not sampled uniformly at random by active learners, $D = .36$, $p = .02$, but no significant difference was found for yoked learners, $D = .22$, $p = .34$. The same pattern was found with the KL divergence in the active group, $KL = .26$ nats (sampled null 95% interval: $[.08, .21$ nats]), and for the yoked group, $KL = .11$ nats (sampled null 95% interval: $[.08,$

¹²It should be noted that the fact that a given distribution at the level of the participant pool does not differ from uniform sampling does not necessarily mean that individual participants sampled randomly. This can also occur, although it is less likely, if participants each were sampling non-uniformly, but were not in agreement.

¹³ 10^4 draws of 6x20 examples, the same size as recorded during the experiment

.22 nats]).

Chapter 4: Tracking others' confidence to select who to learn from

4.1 Introduction

Learning from (or with) others can accelerate the acquisition of knowledge by leaps and bounds, but it also raises a host of novel problems for the learner. In situations of knowledge asymmetry with unknown intent, learners are susceptible to being misinformed by others, and, therefore, have an incentive to develop epistemic vigilance (Sperber et al., 2010). Even in collaborative settings, collaborators will need to decide whether to heed a partner's suggestion, and if so, how much it should be relied upon as a function of the inferred relative expertise. Lastly, in pedagogical settings where the intent is clear but there is inherent knowledge asymmetry, the teacher must assess the reliability of the learner's knowledge to redress departures from the intended teaching goal. Learners may also be faced with the daunting task of choosing a teacher, with sometimes little ability to objectively evaluate teachers given the learner's limited knowledge.

Naturally, a track record of efficacy on the task at hand is of primary importance both for establishing the merits of an advisor/collaborator and monitoring learner achievement. A secondary source of information that can be exploited in addition to accuracy information, or in its absence, is the learner/collaborator's explicit metacognitive judgments or implicit cues to the learner's metacognitive state such as response times, exploration patterns, body language, or speech patterns. This chapter focuses on self-reported metacognitive judgments as they are meaningful and widely used in formal pedagogical and collaborative contexts to which we

would like to generalize, as well as fairly well studied in laboratory settings. However, we acknowledge that there are multiple, more ecological measures that relate to how people report their uncertainty and the cues that people think (rightly or not) that are related to someone's metacognitive state.

On an intuitive level, reported confidence and its covariation with accuracy should be especially useful in predicting the future performance of an agent when a long history of performance on the task is not available or is difficult to surmise and when the subjective difficulty of the task is unknown. Moreover, trial-to-trial variations in confidence can serve as an indicator of accuracy that can improve predictions beyond information about a person's global accuracy level. This leads to the main question of this chapter: **If a partner's confidence is communicated, are humans able to exploit it effectively as a proxy of their competence?**

It is not trivial that humans can and do optimally aggregate their knowledge with that of others in order to improve their own performance. First, an implicit assumption in this line of thought is the broad fidelity of metacognitive expressions since communicated confidence is only useful if it is predictive of performance. Second, since statistically optimal integration entails knowing another person's uncertainty, the question is whether people can, and in fact, do spontaneously monitor cues diagnostic of other people's uncertainty. This can vary from relatively simple tracking of explicit confidence statements or implicit markers (e.g., average reaction times) to more complex monitoring of covariation patterns between accuracy and confidence markers. Third, people would need to aggregate the evidence provided by other agents in proportion to their reliability, forgoing potential egocentric or egalitarian biases. One step further, one of the most compelling questions from our perspective is whether people operate under the previously described fidelity assumption and, therefore, always integrate evidence by weighing it with the expressed confidence, or whether the integration is contingent on inferred (latent) fidelity. While this form of meta-vigilance would have clear benefits for performance, it is unclear whether it is possible to achieve on the fly. In what follows, I will review the pertinent literature for each of these claims.

Before we delve into the first question, it is useful to clearly articulate the quantities that

metacognitive judgments are thought to reflect. A formal distinction has been made in the probabilistic decision making framework between confidence and certainty as different statistical quantities that can be produced on the basis of a decision model. Confidence is computed as the conditional probability of being correct given a specific decision, while uncertainty represents the width of the posterior distribution over the decision variable (Pouget et al., 2016). There is an ongoing debate in the metacognitive literature as to what is being communicated by human participants through experimental self-report secondary judgments (Aitchison et al., 2015; Bahrami et al., 2010a): certainty, confidence, or perhaps a heuristic quantity (vs. an optimally derived quantity on the basis of a decision model). Which quantity is reflected in behavioral measurements is relevant because the upper threshold for the additional benefit that can be derived from receiving a partner's secondary judgments ultimately depends on whether this quantity is an uncorrupted, independent information channel (from the primary decision) or not. In other words, monitoring uncertainty markers in another person is only useful to the extent that those cues are predictive of the internal state of the agent who produced them. Ideally, under the Bayes rational agent assumption, subjective reports of confidence are a function of the objective posterior probability of being correct¹. However, even if the reported confidence only reflects a heuristic (e.g., the magnitude of the sensory stimulus), it can still be useful to an external observer to the extent to which it tracks the internal model of the observed agent.

The assumption of **metacognitive sensitivity** bore out in perceptual decision making experiments as people reported second-order judgments that were well calibrated to performance and objective uncertainty, and correlated with stimulus difficulty and reaction time (Barthelmé and Mamassian, 2009, Kiani et al., 2014, Rahnev et al., 2020). Second-order judgments have also been shown to conform to the predictions of formal models of confidence (Sanders et al., 2016) and potential neural implementations have been proposed (Fiser et al., 2010, Pouget et al., 2016). On the other hand, it appears that changes in the way confidence reports are elicited, such as the simultaneous or sequential nature of first- and second-order decisions, can change whether heuristic or Bayes-optimal computations are used (Aitchison et al., 2015). Further,

¹See Adler and Ma (2017) for a discussion of the expected signatures of Bayesian confidence.

metacognitive abilities (quantified by the type II area under the ROC curve) were shown to be stable within individuals across two perceptual tasks (Song et al., 2011), whereas the same was not true for objective performance. The existence of significant variability in metacognitive abilities within a task, coupled with the consistency of inter-individual differences in metacognition across tasks is a good argument for the usefulness of tracking another's metacognitive judgments as they have generalization potential at least within a restricted task domain (here, 2IFC visual judgments based on orientation vs. contrast).

On the other hand, in more abstract reasoning tasks, there is evidence of the miscalibration of metacognition. Famously, Kruger and Dunning (1999) found overestimation of performance among poor performers and underestimation in high performers, relative to peers. However, the original results of Kruger and Dunning (1999) can be explained without resorting to differences in metacognitive abilities that are performance dependent. Burson et al. (2006) showed that the Dunning-Krueger effect is eliminated by titrating task difficulty and can be best described by a noise-plus-bias model according to which participants are equally poor (noisy) at making metacognitive judgments, and susceptible to task-induced biases: overestimating performance on easy tasks and underestimating it on difficult tasks.

Martí et al. (2018) found that confidence judgments in a Boolean concept learning task (with a very large hypothesis space) were best predicted by local or global accuracy rather than any quantity derived from an ideal learner model (e.g., model uncertainty over the hypotheses, the probability of the best hypothesis), which added little additional value in terms of predictability. However, it is not surprising that, in the presence of feedback, local accuracy would serve as an effective and salient cue for confidence judgments. The differences in variance explained (R^2) between the models using model-based vs. behavioral-based predictors were not large, but meaningful (maximum 5-10%). The ideal model explained somewhere around 50% in the variability in accuracy, so a potential counterargument is that the participants might have followed a different (perhaps resource-rational or heuristic) model to solve the task (and thus the original comparison is not useful).

More convincingly, in a follow-up experiment in which feedback was removed (and pre-

senting only one trial of the task), the variance accounted for by model-derived quantities did not increase. It should be noted that the Martí et al. (2018) model did not fit the accuracy of the participants for this follow-up experiment, either due to participants not actually learning the task or due to data sparsity. Assuming that the model accurately characterizes performance in this second experiment, the result implies that model-derived measures were not used or computed to provide certainty judgments, despite the fact that participants had the quantities needed and the ability to do so. In essence, the study suggests a dissociation of the factors that drive learning decision-making from those that determine certainty, which could significantly limit the usefulness of metacognitive judgments for inferring another person's uncertainty.

To sum up, while there are clear instances where metacognitive assessments faithfully convey model-based quantities that would be highly informative, changes in the way in which confidence reports are elicited, as well as the nature and complexity of the task, can easily result in a switch towards a heuristic strategy. In cases where the reported confidence does not reflect a model-based quantity, it is at worst redundant (still correlates well with accuracy and difficulty, as is the case in the Martí et al. (2018) study), and therefore, there are smaller gains and losses to be made from additionally monitoring another's confidence. It remains unclear, however, in cases where trial-by-trial confidence reports are optimally derived, how much they would empirically benefit a potential observer who is attempting to make inferences about the internal model of another. This would be an interesting theoretical and empirical direction of study.

The second claim to be assessed is that humans **spontaneously monitor cues to another's certainty**: explicitly communicated confidence or implicit cues such as response times and non-verbal communication². Sensitivity to others' metacognitive states would be evidenced by explicit differential inferences about the competence or trustworthiness of social partners who are matched for accuracy, but diverge in their confidence. In fact, even young children are sensitive to the confidence of others, as shown by their increased willingness to learn from confident informants (Birch et al., 2010). In addition, we can ask whether humans track (and

²The list is not exhaustive. For instance, Pulford et al. (2018) found a good signal for who is more confident in a dyad was who spoke first.

expect) that the confidence of others positively correlates with their accuracy and negatively relates to task difficulty.

The main line of inquiry in this direction tackled the undue influence of confidence on the perceived credibility of a witness (Cutler et al., 1988) or financial advisor (Price & Stone, 2004), both situations of clear ecological relevance. Results have been explained within the game-theoretic framework through the use of the confidence heuristic (Pulford et al., 2018; Thomas & McFadyen, 1995), whereby people assume that the stated confidence is a stand-in for the informational reliability of a source. Naturally, this opens the door to strategic manipulations of confidence, as overstating confidence can benefit agents by making them appear more competent and accrue more influence and social status³. On the flip side, this means that people also need to develop mechanisms for epistemic vigilance to defend themselves from such bad faith actors (Sperber et al., 2010).

Subsequent experimental work has shown that while people are sometimes vulnerable to the (intentional or unintentional) overconfidence of others, with additional explicit evidence as to the unreliability of the overconfident agent, people can overcome biases favoring overconfident others. For instance, in the context of a mock trial with participants acting as jurors, Tenney et al. (2007) manipulated the perceived accuracy and confidence of eyewitnesses in collateral statements that were peripheral to the central issues of the court trial. Results revealed an interaction between the accuracy and confidence of eyewitnesses as participants perceived highly confident witnesses as more credible than unconfident ones when they were correct (or in the absence of knowledge about the veridity of their testimony), while the opposite pattern was observed for inaccurate eyewitnesses. Inaccurate but overconfident eyewitnesses were especially penalized for their miscalibration, underscoring the importance of competency information. Based on their findings, Tenney et al. (2007) proposed an extension to the confidence heuristic, arguing that there is a default assumption that stated confidence is a good predictor for accu-

³Johnson and Fowler (2011) have gone so far as to argue that overconfidence, even just as poor metacognition, so self-deception cf. as strategic deception of others, is optimal (maximizes individual fitness) given competition under uncertain and asymmetric payoff structures. They use a restricted model of decision making in the competition over resources in which agents decide whether to claim a resource only as a function of their and their competitor's perceived likelihood of winning a potential fight. Simulation studies of this model showed that a stable bias toward overconfidence emerges at the population level.

racy only in the absence of external evidence to the contrary. That is, unless proven otherwise, people will believe a confident informant is an accurate one.

Similar findings come from a study by Sah et al. (2013) in which they directly tested the calibration assumption by modulating the availability of accuracy information. Participants were asked to guess the weight of a person based on a photograph of their face by assigning probabilities to several possible weight ranges. The participant then received advice from one of four advisors in the form of an estimate and an associated continuous confidence rating. Participants were then allowed to adjust their initial rating. The four advisors were generated using a between-participants 2 x 2 factorial design corresponding to low and high accuracy and confidence. Accuracy was manipulated by showing feedback according to which the advisor was consistently correct or incorrect (100% vs 0% accuracy) in five successive trials at the beginning of the task. Confident advisors had confidence ratings of 95% on average, and unconfident advisors had ratings of 30% on average (the confidence scale is not restricted to 50-100 because participants stated confidence in one of several weight ranges).

Sah et al. (2013) found the expected interaction between these factors on explicit questions about advisor credibility at the end of the task. Namely, among the two accurate advisors, the more confident one was more credible, whereas the unconfident inaccurate advisor was deemed more credible than the confident inaccurate advisor. As predicted by the calibration hypothesis, in a follow-up where there was no feedback, participants' credibility judgments were driven solely by the confidence of advisors. Interestingly, the results did not show an effect of the calibration profile on how much participants relied on the advice they received, even when performance-based monetary incentives were added. The null results may be explained by the fact that there was no (or very limited) exposure to the covariance of advisers' confidence and accuracy. Intuitively, if an advisor is 100% correct (or incorrect), there is no benefit in tracking confidence or modulating decisions based on it in a trial-by-trial fashion. Similarly, participants could have assumed that advisers' confidence was not modulated by accuracy but instead reflected something akin to a 'personality trait', so there was little scope for relying on it to improve performance. The less likely (and disquieting) alternative explanation would be

that miscalibration affects the credibility of informants, without any consequences in terms of discounting their advice in future interactions.

The match between confidence and accuracy certainly influences the perceived credibility of others. However, based on the literature presented thus far, it is not clear how refined this ability is beyond cases of flagrant misrepresentation of one's knowledge. In the studies discussed so far, the agents judged by participants were either entirely accurate or inaccurate in their largely one-shot advice, and their confidence was also discretized or at least highly contrasting as in the case of Sah et al. (2013). This made the calibration profile a clear-cut, binary dimension. A more substantial test for calibration would assess whether human preferences for calibrated others develop across repeated experiences with others, and are quantitatively related to the statistical relationship between confidence and accuracy. To this end, it is important to be more explicit about what calibration entails statistically.

Yates et al. (1996) discuss calibration within the more general context of what makes a good probabilistic forecaster. Specifically, they are interested in preferences for advisers who make repeated continuous probability predictions for the occurrence of a binary target event⁴. There are two statistical properties of probabilistic predictions that are often confounded. The first is the discriminability, namely, whether predictions or, in our case, confidence ratings are different for correct versus incorrect decisions. Second, judgments are calibrated if there is a close to one-to-one match (high correlation) between the binned, properly scaled confidence ratings and the probabilities of being correct. In the context of forecasting, it is possible to have two agents with the same predictive error, but different trade-offs between discriminability and calibration. Discriminability is certainly a primary prerequisite for confidence to be a useful signal for partners and supersedes calibration. It is possible to re-calibrate the advice received, but an unpredictable signal is useless. For example, once you learn that a cooperative partner is only 70% accurate when they state they are maximally confident, you can redress your expectations of their performance accordingly. If, on the other hand, there is no statistical dependency

⁴While these are not traditional, explicit post-hoc confidence judgments, it is possible to translate such probabilistic forecasts into traditional confidence statements by first computing binary decisions (depending on whether the estimated probability was below or above 50%) and expressing the probability of that decision being correct (within the .5 - 1 range).

between confidence and accuracy, there is no information to gain. Yates et al. (1996) designed forecasters who had equal quadratic error, but one was better calibrated, the other was more discriminative (see Figure 4.17 in the Supplementary Information to visualize forecaster differences). Results showed a preference for the forecaster with the higher discriminability, who was preferred over the more calibrated one. However, the more discriminative agent was also by design more extreme in their probability judgments, which was a potential confounder.

Following the conceptual approach of Yates et al. (1996), Price and Stone (2004) set out to specifically test whether people prefer advisors who produce more extreme judgments of confidence, particularly overconfidence, when discriminability and overall accuracy are controlled for. In the Price and Stone (2004) design, two financial advisors produced multiple sequential binary decisions about whether different stocks would increase or decrease in value alongside an estimated likelihood of that event expressed in percentage points. At the same time, participants were told whether the stock's value had increased or not. The two financial advisers made 24 such forecasts each (separately), following which the participants were asked whom they would choose to hire. The agents were equally accurate (75%) in predicting the increase/decrease of stock prices. However, the likelihood ratings of the extreme agent were exactly 15 points more extreme than those of the calibrated agent (average likelihood judgments: 71.67% and 86.67%). Importantly, since judgments above 50% in likelihood are interpreted as categorical judgments that the target event will occur, the binary predictions of the two agents were equivalent and, therefore, the ability of the likelihood ratings to discriminate correct predictions was matched.⁵ The results supported the confidence heuristic as more participants preferred an overconfident advisor in lieu of a more calibrated advisor with the same accuracy (moderate effects replicated across three experiments). Moreover, participants also viewed the advisor with more extreme confidence judgments as more knowledgeable (but not more honest or optimistic). In a further replication experiment, the accuracy of the overconfident forecaster

⁵The discriminability of the likelihoods for the two target events (the slope dimension as described by Yates et al. (1996)) was steeper for the more extreme agent. This is an unavoidable confound given the current design, but one that was eliminated in the third experiment of Price and Stone (2004) in which participants expressed confidence for their binary decisions within the 50-100% interval. Results were replicated, although with a smaller effect size (63% preference for the extreme advisor, BF_{10} calculated for this result 1.82 in favour of the alternative).

was overestimated while the moderate forecaster was considered less accurate⁶.

Although the Price and Stone (2004) design allowed the quantification of a learned relationship between accuracy and confidence, it can be argued that in the case of continuous confidence assessments for binary outcomes that are driven by a latent probability, confidence is not strictly perceived as a second-order decision, such that being 90% confident in the correct prediction of an event can be construed as being more correct than being 70% correct in that decision. This could then lead participants to the heuristic of preferring the agent who is "more correct" more often. This would indeed be the overconfident agent. And, in fact, this is what Price and Stone (2004) have found in the accuracy ratings of the participants. These ratings were made after participants chose the adviser who they wished to hire, which could have also led to a retrospective bias in an attempt to justify their choice (although the manipulation did not affect the ratings of advisor honesty and optimism). The alternative explanation could be that unlike in the Tenney et al. (2007) and Sah et al. (2013) experiments discussed above, here the miscalibrated advisers did not clearly violate the a priori assumption that confidence is a good proxy for accuracy and still benefited from the confidence heuristic.

Lastly, so far only monitoring of verbal or numerical confidence judgments was discussed, but there is evidence that people will spontaneously monitor implicit cues to the confidence of others. Notably, Patel et al. (2012) asked participants in turn to perform a contrast discrimination task requiring a motor response and an associated confidence judgement themselves, and then to observe a demonstrator (one of two) perform the same task. Participants then made predictions about the trial-by-trial confidence of the demonstrator without knowing the identity of the stimulus they were responding to. They were able to make fairly accurate predictions by using the trial-by-trial variations in response times (with shorter RTs being indicative of higher confidence). Of course, this linear trend between RTs and confidence is present not just within individuals, but also across individuals, so it would have been useful to replicate these findings using two demonstrators (within participant) which would allow to test the quality of individualized predictions against shuffled predictions to test the degree to which individuals'

⁶Price and Stone (2004) also showed that the preference for the overconfident agent was significantly positively related to right-wing authoritarianism beliefs.

idiosyncrasies are captured. However, more interestingly, Patel et al. (2012) showed that the linear relationship between RTs and predicted confidence of the demonstrator was modulated by the participant's (intra-individual) relationship between response time and confidence. Particularly, if they rated the demonstrator to be more confident on average than themselves, then they were also more likely to be slower than the demonstrator. The slopes describing the relationship between participants' own RT-confidence mappings correlated with those describing the slopes of the predicted RT-confidence mapping for the demonstrator. Again, it would have been interesting to see the extent to which one's extemporaneous predictions for oneself differ from predictions for another empirically. It is indeed unclear whether in artificial situations where people have to make explicit predictions for behaviors that normally are considered automatic, they can use additional implicit self-knowledge or whether the same model would be used for the self as for another person.

All in all, people monitor cues to informants' confidence spontaneously, as well as the calibration of confidence to accuracy agents in situations where this is straightforward to do. It is less clear what happens in situations where tracking the covariance of accuracy and confidence across time has higher cognitive costs, which can prompt the use of the confidence heuristic. Indeed, Tenney et al. (2011) have shown that children and adults who are cognitively taxed (by performing a secondary task) do not make use of calibration information adequately.

Moving on, the final assumption is that people can **incorporate information from others appropriately as a function of their metacognitive judgments** in order to improve decision making. A burgeoning literature on perceptual decision making under uncertainty is showing that humans can integrate different sources of environmental evidence near-optimally (e.g., Ernst and Banks, 2002). More recently, there has been considerable interest in extending this framework to joint decision making, asking whether people can benefit from the knowledge of their social partners in a similar manner by utilizing information about their uncertainty.

Notably, Sorkin et al. (2001) tackled the issue of information integration in groups of collaborating individuals by drawing direct parallels to the multisensory cue integration literature. They proposed that what group members are communicating with each other is the distribution

of the decision variable. To be more concrete, in a group signal detection task, individuals would independently communicate the mean of the perceived Gaussian signal and its variance (of course, not numerically, but implicitly by freely communicating with each other). Given these two pieces of information, from a theoretical standpoint, the ideal group's collective decision making should always supersede the performance of any individual in the group. This is precisely what the experiment showed, as the groups had higher detection performance (d') compared to any of the individuals constituting the groups, and performance was close to that of their ideal model. Moreover, they showed that, as predicted, group performance increased with group size, decreased with correlated judgments, and more competent group members were more influential.

Bahrami et al. (2010b), however, proposed that what collaborators were expressing through their confidence was the ratio of the mean signal to its standard deviation. This is a meaningful deviation from the Sorkin et al. (2001) account, as it opens the door to explaining inefficiencies in group decision making on the basis of a formal model that departs from the Bayesian optimal cue integration. While the two group decision models made similar predictions for collaborators with similar sensitivities, they can be dissociated when their sensitivities diverge. In particular, large differences in individual sensitivities would result in poorer group performance, as the expected collective benefit is a linear function of the ratio of two partners' sensitivities. Across a series of experiments conceptually mirroring Sorkin et al. (2001), Bahrami et al. (2010b) tested the predictions of their model (Weighted Confidence Sharing, WCS) with a 2IFC perceptual discrimination task in which participants performed both individually and in dyads. In the joint condition, collaborators first made individual judgments and, if there was disagreement, they were allowed to communicate with their partner freely. One of the participants were randomly chosen to make the joint decision. Feedback was then offered. Experiment 1 replicated previous findings of a higher group sensitivity (quantified by a higher psychometric slope) than that of the best individual performer. In Experiment 2, additional noise was added on some trials for dyad members either symmetrically or asymmetrically in order to experimentally introduce individual differences in sensitivity. In the unequal noise con-

ditions, group performance dropped below that of the best individual member's performance. This is in direct disagreement with the original Sorkin et al. (2001) model prediction, but in line with the predictions of the WCS model.

Bahrami et al. (2010b) further explored whether the success of the dyads is attributable to the fact that they could communicate about their confidence or whether they relied on feedback. One can imagine that, given enough time, participants could infer the average reliability of their partner just based on observing the other's decisions alongside the group feedback. To dissociate these different explanations, two additional experiments were conducted: one in which feedback was withdrawn, to test whether communication alone could produce the previously observed effect, and one in which dyad members were not allowed to communicate, testing whether feedback was sufficient to elicit the effect. Results indicated that feedback was not necessary to obtain the superior group performance, but communication was. It is perhaps not particularly surprising that participants were not able to capitalize on the presence of feedback to estimate their partner's reliability given the trial-by-trial random noise addition. Perhaps a fairer test of this hypothesis would involve a global (or blocked) manipulation of the sensitivity of the collaborative participant. Bahrami et al. (2012) used such a design to show that nonverbal displays of confidence (numerical ratings) were less conducive to "counterproductive collaboration" than verbal communication. This indicates that participants were better able to make inferences about reliability when confidence was unambiguously presented and thus avoided the full integration that would lead to decrements in performance. However, confidence ratings were always presented in this experiment.

Thus, there is clear evidence of incorporation of information from others on the basis of uncertainty communication, even if sometimes it leads to deleterious outcomes. However, previous experiments relied on the assumption of good metacognitive calibration for both agents, since the predictions were based only on type I performance (their psychometric curves), assuming that participants were able to perfectly communicate their uncertainty. This is problematic, especially considering that people have been shown to vary greatly in their metacognitive abilities (Fleming et al., 2010). In a follow-up experiment aimed at addressing this concern,

Pescetelli et al. (2016) tried to disentangle the effects of sensory evidence and social factors on the integration of information. They fixed the accuracy of dyad members by using a staircase method to keep performance at threshold and ensure that global accuracy would not be diagnostic of their partner's reliability (first-order and second-order performance were independent). The correlation between the mean metacognitive sensitivity of the dyad (measured by A''_{ROC}) and the collective performance of the dyad (relative to individual performance) was positively and significantly correlated. This shows not only that previous assumptions were flawed, but that the participants were also operating under the same assumption of equal metacognitive sensitivity. Indeed, while the trial-by-trial correlation between accuracy and confidence is responsible for the collective dyad benefit, participants were not able to go beyond assigning higher influence to the most confident member, which resulted in poor performance when that reported confidence was not well calibrated to accuracy (Pescetelli et al., 2016).

The Pescetelli et al. (2016) results resonate with Mahmoodi et al. (2015), who used a very similar experimental design to show that (across cultures) people use more or less equal weighting of group members' decisions despite differences in competence. They implemented a Bayesian reinforcement learning model (Behrens et al., 2008) that estimates the probability of a correct decision for an agent given the history of their decision accuracy, and then weights this probability by the agent's reported confidence. The model assumes that an individual's choice for the dyad will then be a linear combination of these probabilities with a fitted weight. The less sensitive member of the dyad, who was less metacognitively calibrated, tended to underweight their more competent partner and the more sensitive member overweighted their less competent partner, essentially falling into the trap of the Dunning-Kruger effect (although also compatible with the Burson et al. (2006) interpretation of the effect). In a series of control experiments, Mahmoodi et al. (2015) showed that this effect survives when larger, more salient differences in sensitivity were introduced (by experimentally making the task harder for one of the participants), when reminders of past performance were given (alleviating memory load), and when adding monetary incentives.

Additional evidence has been consolidating to show that groups of participants are subop-

timal when aggregating information under conditions of competence asymmetry. Bang et al. (2017) explored whether group members can align their confidence expressions and recalibrate them to a common metric, which would then assure the optimality of their joint decisions. They propose that since inferring another agent's function to link probability correct to their reported confidence is computationally difficult, people use a heuristic strategy instead. Specifically, they suggest that in social decision making, people will match their reported confidence with each other (e.g., they will match mean confidence or the distribution of confidence). Intuitively, when collaborators have similar accuracy, confidence matching leads to optimal dyad performance. On the other hand, when the two members differ in competence, there are two opposing patterns. For calibrated dyads in which the more accurate collaborator is also the more confident, confidence matching will lead to costs in dyad performance. For miscalibrated dyads in which the less accurate member is the more confident one, confidence matching will improve dyad performance. The results revealed markers of confidence matching: in the social group decision making condition, the differences in mean reported confidence between partners were smaller, and the confidence of dyad members was significantly correlated only when performing the task in a group. Furthermore, in line with predictions based on the heuristic, dyads' departure from optimality correlated positively with an increase in the dissimilarity of the members. In dyads comprising one participant and one simulated collaborator used to create calibrated and miscalibrated dyads, on average, dyad accuracy increased relative to the initial individual accuracy for poorly calibrated dyads and decreased for calibrated ones.

As described above, a rich body of work bore out of the idea that human communication enables cooperation by incorporating information optimally leading to both positive and negative consequences for joint outcomes. Communication of confidence has been shown to boost dyad performance (generating the wisdom of the crowd effect) even in the absence of feedback. At the same time, participants were not able to adjust in situations with asymmetric competence and miscalibration of confidence with accuracy. Such situations are by no means rare outside of the lab. If anything, one would expect that more abstract cognitive tasks offer more chances for miscalibration and asymmetric competence than highly controlled psychophysics tasks.

The biases observed in these experiments were largely failures to either recognize correctly or account for the relative abilities of a collaborator, assuming instead that they were more or less equal. This can emerge either from not having an appropriate model of the other person and bootstrapping from our own/using a heuristic, having a strong prior of equal ability, or it can be an audience effect. Specifically, due to the social nature of the repeated collaboration with a partner (no rewards and potential social costs), equal consideration of the partner's input is a sensible strategy. It would be interesting to check if the same effects are observed if the joint choice made by the decision maker was not shown to the collaborator.

The following section will reiterate some of the gaps in the current literature and propose how some of them can begin to be addressed.

4.2 Study motivation

As outlined in the previous section, the relationship between one's confidence and accuracy certainly influences one's perceived credibility and sway on others (Tenney et al., 2007; Tenney et al., 2008). However, based on the literature presented thus far, it is not clear how refined this ability is beyond cases of flagrant misrepresentation of one's knowledge.

The evidence for sensitivity towards the metacognitive skills of informants comes from experiments that either test explicitly the assumption that a confident agent should also be correct in a one-shot judgement, or compare informants who are perfectly calibrated with informants who are particularly poorly calibrated. Specifically, in the Sah et al. (2013) experiment, across several trials, an adviser was either always accurate or always inaccurate and used only the extremes of the confidence scale (average of 30% confidence for the low confidence condition and 95% for the high confidence condition respectively). There was a significant interaction between accuracy and confidence in the explicit rating of advisor knowledgeability and trustworthiness, which is indicative of a calibration preference, and yet participants did not modulate how they revised their judgments based on the calibration profile of the advisers. This is not surprising given that the discrepancy between the calibrated and miscalibrated agents was easy

to spot explicitly, but there was no extra benefit that calibration information could provide for the integration of information. Regardless of the expressed confidence, participants ought to use the advice of the always correct agent and discard the advice of the always incorrect agent.

On the other hand, in an experimental setup in which the discriminability of the informants contrasted was matched and calibration differences were more graded and had to be learned through repeated experience, Price and Stone (2004) instead found a preference for an overconfident advisor over a more calibrated advisor with the same accuracy. There are several potential reasons for the discrepancy in these results. First, it is possible that only very large differences in calibration, such as those introduced by Sah et al. (2013), are perceptible. However, given the large body of evidence on human unsupervised statistical learning abilities (Fiser & Aslin, 2001), this explanation seems unlikely.

Second, Price and Stone (2004)'s findings in support of the confidence heuristic could be the result of the specific format of the forecasts which lead to the overconfident advisor being (erroneously) judged as more accurate. For instance, a forecast that a stock price has a 90% probability to increase can be judged as more accurate than a 70% forecast, even though both forecasts make the same correct prediction (both are above 50%). Thus, if participants were just counting who was more correct more often, rather than computing the average probability score over all trials, they would end up with the conclusion that the overconfident forecaster was more correct which could then influence preference judgments.

Lastly, the cover story employed by Price and Stone (2004), a comparison between two male financial advisors competing for a job (depicted in business suits and briefcases), may have elicited specific stereotyped expectations of people in such roles. Indeed, the authors found that participants higher in personality measures of authoritarianism also showed higher preferences for the overconfident adviser. Indeed, it would be interesting to check whether at the population level, there is more vigilance about the overconfidence of financial forecasters now, given the the financial crisis that occurred since the article was published.

In the proposed experiments, we aim to reconcile the literature and test the sensitivity to (and preferences for) calibration using designs in which advisers exhibit graded, quantifiable

differences in the statistical association across time between accuracy and confidence. We share Yates et al. (1996)'s opinion that that a key basis for preference judgments about advisers is the ability of confidence judgments to predict accuracy. Therefore, it is crucial to manipulate both the discriminability of the advisers' confidence and their calibration to be able to tease apart their contributions.

First, in light of the results of Price and Stone (2004), the preference for a more discriminative advisor in the original Yates et al. (1996) study needs to be further replicated to ensure that they hold when the extremity of the forecasts is controlled for. Second, in the absence of differences in predictability, as supported by previous experimental evidence, the confidence heuristic could lead to a preference for overconfident agents. Alternatively, overconfidence may be penalized given that overconfident agents can be perceived as using their confidence judgments for deceptive purposes. A third option is that there is a general preference for calibrated collaborators, compared to both over- and under-confident others, even in a situation of matched predictability. Calibration speaks to the self-awareness of others and could suggest that they have a good understanding of the advice they are giving, which should increase preference for a social partner generally (beyond evaluations of whether to trust specific assertions). More cynically, it is less mentally effortful to interact with calibrated agents whose confidence can be used at face value (i.e., to whom the confidence heuristic can be applied).

In light of this, we proposed a study with a series of experimental manipulations starting from a conceptual replication and extension of the Price and Stone (2004) experiment. The goal of the first experiment was to assess whether preferences for future collaborators depend on previous inferences about the calibration profile of the agents who were equally accurate, their confidence differences, as well as the extent to which their confidence judgments were predictive of their accuracy, and dissociate between these three factors.

Price and Stone (2004) also found a tendency for men to prefer more extreme advisers to a greater extent than women, so the experiment will fully counterbalance the gender of the participants and that of the avatars.

The second study intended to go beyond testing general preferences for collaborating with

one agent versus the other, which may be impacted by a multitude of intertwined factors (long terms chances of gains through cooperation, perception of being honest or manipulative etc.) to test how knowledge about the metacognitive skills of others could be leveraged in practice to improve individual performance.

Specifically, we aimed to contrast participants' adviser choices with optimal decisions assuming that participants monitored the functional mapping between confidence and accuracy. Given sufficient experience with advisers, people should be able to build a simple model that allows them to predict the probability of an adviser being correct given the stated confidence of an agent and any other potentially relevant factors (e.g. task domain, audience). This prediction would constitute an inferred, recalibrated confidence. Only a metacognitively sensitive, calibrated agent would have equivalent explicit and recalibrated confidence.

Therefore, the second study directly tested whether participants were able to adjust, in other words, to recalibrate, the confidence of advisers before using it to make individual decisions. Previous results would cast doubt on this ability. The already discussed Sah et al. (2013) results and (Mahmoodi et al., 2015) could be taken to suggest that miscalibration affects the credibility of informants as explicitly reported, without any implicit consequences for behavior in terms of discounting their advice in future interactions. We have already argued that in the case of Sah et al. (2013), there was little benefit in using calibration when incorporating advice due to the floor/ceiling accuracy of advisers. Moreover, we believe that metacognitive differences may be exploited in a less cognitively demanding, but more abstract task (c.f. (Mahmoodi et al., 2015)). It is possible that participants might have stronger priors of equal expertise and metacognitive sensitivity in the case of perceptual decision making compared to tasks involving more abstract cognitive skills. Indeed, according to work discussed earlier (Martí et al., 2018), this prior would track the ground truth. It would be informative to have measures of for instance the Dunning-Krueger effect across different task domains (and not just task difficulty levels).

A further critical deviation in our proposed experiment is that participants will not be engaged in a joint task, rather they will be tasked with the evaluation of two potential advisers.

We wanted to explore whether the equality biases observed by (Mahmoodi et al., 2015) are purely egocentric and relate to either the increased demands of closed loop interaction or social norms around it, and whether they can be avoided when participants need to integrate evidence coming from two other people.

4.3 Experiment 1

4.3.1 Introduction

The first experiment was a conceptual replication and extension of Price and Stone (2004), aiming to test preferences for collaborating with advisors who differ in their discriminability and/or calibration. Distinctly from Price and Stone (2004), the predictive ability of confidence was manipulated alongside the marginal match between the probability of being correct and the probability of being confident. Calibrated agents were compared to both under- and overconfident agents, to assess whether departures from calibration have a symmetric effect.

Participants observed two agents simultaneously performing a novel categorization task which allowed them to monitor the covariation between the agents' binary confidence expressions and accuracy. Following this observation phase, participants chose one of the agents as a future collaborator for a following task. Categorization was preferred to the financial advisor task to remove any prior stereotyped expectations for the confidence of such agents. Crucially, the agents were correct the same number of times, but their calibration and discriminability ability was experimentally manipulated across eight between participants conditions (see Manipulation section below).

In planned follow-up control conditions, we assessed the perceived accuracy of the observed advisors and no preference judgments were made by participants. This was necessary to ensure that any above-chance preferences in the experimental conditions were due to the experimental manipulations and not to the mistaken perception that one of the agents is more accurate than the other⁷. A within-participant measure of the difference in perceived accuracy

⁷Another control was planned contingent on a null result in partner selection preferences to test whether

(like in Price and Stone (2004)) could be biased by the preference decision as participants could assign a higher accuracy to their preferred agent to justify their choice (since it is likely that the calibration will not be an explicit, but that implicit, reason for their choice).

A second sanity check task was used to ensure that the statistics of the two advice sources could be accurately perceived by participants. This control was run post-hoc to explain the unexpected finding that the condition in which a calibrated agent was contrasted to an agent whose confidence was independent (and therefore unpredictable) of confidence. In order to test this, we explicitly asked participants to report back marginal and conditional probabilities in the context of a task that was equivalent to the main experiment in all aspects except the cover story. A non-social cover story was used to eliminate any effects due to prior expectations about statistical associations between the variables presented. This control was applied to the aforementioned conditions alongside a condition in which we observed the strongest preference effects to check whether similar preferences would emerge.

Quantifying metacognitive performance on the experimental task

It is necessary to clarify at this point what is meant by calibration in the context of this task. A calibrated agent has been canonically defined as being confident to the extent that they are accurate. For the current task, this would entail making the same proportion of 'confident' judgments as correct decisions. Some conditions contrasted a calibrated agent with the marginal probability of being confident equal to the probability of being correct, to another agent who had a higher (overconfident) or lower (underconfident) probability of being confident. We will use the terms over- and under-confident to refer to agents who have different marginal probabilities of being correct relative to the probability of being correct.

We have argued that the predictive ability of confidence is a more relevant factor for partner selection than a match in the marginal probabilities of being confident and correct. In order to quantify the informativeness of the agents' confidence, the mutual information between

participants accurately perceived the difference in the marginal probability of being confident. This control was not conducted.

confidence and accuracy was computed for every agent⁸. Mutual information (also known as information gain) quantifies how much of the uncertainty about accuracy of the agent has been reduced by knowing their confidence. The equation below describes the mutual information between accuracy (Y) and confidence judgments (X), where \mathbb{H} denotes entropy:

$$\mathbf{I}(A;C) = \mathbb{H}(A) - \mathbb{H}(A|C) = \mathbb{H}(A) - \sum_{c_i \in \mathcal{C}} p(\mathcal{C} = c_i) \cdot \mathbb{H}(A|\mathcal{C} = c_i). \quad (4.1)$$

If confidence and accuracy are independent, the information gain is null. On the other hand, if the two have a perfect one to one correspondence, the information gain is maximal. The marginal level of confidence will constrain the extent to which the covariation of confidence to accuracy can be used to increase the informativeness of confidence. Further, given a marginal probability of being confident (and the fixed accuracy), it is possible to derive which conditional probabilities will lead to the greatest accuracy level.

There are two possible ways in which agents who produce the same decisions can differ in their binary confidence judgments. First, they can both produce the same Bayes-normative continuous confidence judgments, and then set different thresholds to binarize continuous confidence when reporting it. The agent who uses a lower threshold would be more confident overall than their counterpart. That is, they will produce more 'high confidence' judgments in general, meaning that they will be relatively less likely to be correct when stating they are confident, and also when they say they are not (but not necessarily to the same extent).

The other way in which two otherwise Bayes-rational agents can reach different confidence levels for the same decisions is if they assume different generative models for the data they observe. For instance, we can imagine a simple categorisation task where stimuli corrupted by noise come from categories defined by two equal variance Gaussians symmetric around zero. Decisions on the category of the stimuli are deterministic as a function of whether the perceived stimulus is lower or higher than zero, so independent of the perceived spread of the two cat-

⁸Alternatively, a chi-square test of independence could have been used, but mutual information provided a more intuitive interpretation.

egories. However, if one of the agents computes the confidence judgments believing that the variance of the two categories is larger, then they will produce comparatively underconfident judgments. Similarly, if the agent believes in a very narrow category, their resulting confidence judgments will be higher than those of their equally accurate counterpart. Supplementary Information 4.6.1 details the generation of confidence judgments for different agent profiles by showing simulated data.

4.3.2 Methods

Participants

759 participants (377 female, 376 male, 4 nonbinary) were recruited through the Prolific and Testable Minds online testing platforms. The average age of the participants was 30.70 years (SD = 10.88). All participants were fluent English speakers of various nationalities, predominantly located in Europe and North America.

Participants were paid 1.6\$ for the completion of the study and were awarded a bonus of .5\$. The bonus was used to incentivize participants to pay attention to the observation phase of the experiment as they were informed that the bonus would be performance-based. However, all participants were awarded the bonus upon completion of the experiment.

The recruitment goal was 64 participants for each of the main experiment conditions and the two nonsocial control conditions and 56 for the two planned accuracy perception control conditions.

This sample size was determined such that an a preference of around .65⁹ could be distinguished from chance with power .8 in three primary planned FWER corrected binomial tests conducted on subsets of the conditions to test whether participants exhibited: a preference for a more informative agent when calibration is controlled; there is a preference for a calibrated agent when compared to an overconfident agent, and when compared to a underconfident agent. In the case of a null result, this sample size would be sufficient to provide a moderate to strong

⁹The proportion of preferences for the more extreme agent varied across the Price and Stone (2004) experiments from .63 to .71.

Bayes Factor in favour of the null. Further, the sample size for the accuracy perception control was planned to allow for one sample t-test per condition to uncover a difference of 3-5 points (on the 0-100 scale) with a standard deviation of the sampled population around double the difference.

One participant was excluded from the analysis for not meeting the required age of consent (minor) and another participant was excluded for noncompliance with the experiment instructions.

Ethical approval was obtained from the United Ethical Review Committee for Research in Psychology (EPKEB) in Hungary.

Tasks

Main task

The experiment started with an observation phase. Participants were introduced to avatars representing two other participants and were told that these participants were previously taught how to categorize four amorphous shapes in two distinct categories (e.g. "wug" or "dax"). The participants' task was to carefully observe as these participants were tested on the categorization task that they would themselves have to perform later. It was stressed that the agents were not communicating, did not see each other, and did not receive any feedback. The reason for this was to prevent participants from forming the expectation that the agents' performance should improve with time, or that they were playing against each other.

In each trial, a shape first appeared on the screen for 1000 ms. The true category of the shape, the two agents' binary confidence ("High" or "Low"), and their proposed categories were revealed sequentially every 1750 ms. Each trial was followed by a 750 ms inter-trial-interval. Each shape was presented 15 times, amounting to 60 observation trials in total.

The schematic avatars corresponding to the two agents could be distinguished by clothing and hair color, but were identical in all other respects. Participants were always presented with two avatars of the same gender to avoid any gender-based preferences. Within a condition, a balanced factorized assignment was used for the the gender of the avatars and that of the

participants. For some conditions, there are small departures from the gender balance since the self-reported gender of the participants did not match that recorded by the online recruitment platform. The self-reported gender is used for descriptives and all analyses.

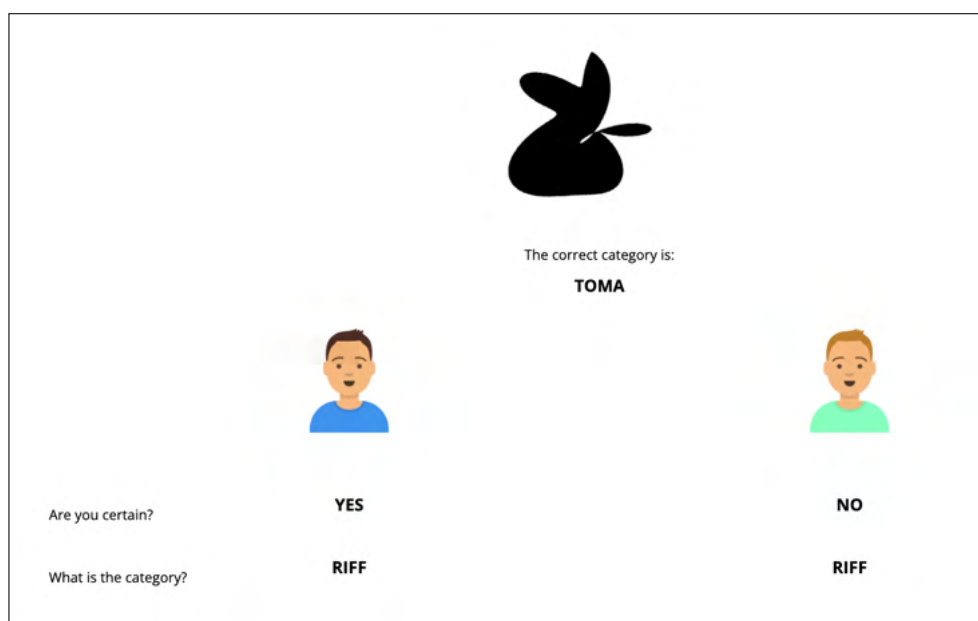
The identity and positioning of avatars on either side of the screen was randomized. Category names and shape sets were also randomly chosen from three possible options. Agents responded with the two category labels an equal number of times and were matched in their performance within each category.

Following the observation phase of the experiment, participants were asked to choose which of the two agents they wished to collaborate with in the future on a similar categorization task. They were reminded that the bonus was contingent on good performance on this coming task. After the forced choice, participants could offer an open-ended justification for their selection. Following this, they rated the accuracy of the selected agent on a scale from 0-100 (corresponding to never - always correct).

Lastly, the participants completed a short 12 trial version of the categorization task they observed. Performance on this task was used as an attention check for the observation phase. Participants were then debriefed before concluding the experiment.

The experiment was completed on average in 13 minutes and at most within of 45-50min, which was the maximum completion time allowed by the recruitment platforms.

Figure 4.1: Example trial from the observation phase.



Planned control conditions: Perception of relative accuracy

The observation phase of the experiment was identical to that in the main task. The key difference was that the participants were asked immediately after the observation phase to rate the accuracy of both agents on a 0-100 point scale. Then they answered the 2AFC preference question as in the main task, followed by the open-ended question. Accuracy ratings were performed first (unlike in the main task) to avoid the potential bias that could stem from a desire to be self-consistent with a previously expressed preference or the halo effect. Participants then rated the overall confidence of the two agents, also on a scale from 0-100. Lastly, they completed the short categorization task.

Post hoc control conditions: Perceptibly of statistical differences

In the nonsocial control task, the statistical structure and procedure of the observation phase was kept intact, but the cover story was replaced. Participants were told that they were going to observe two translation apps (distinguished by the color of their logos) as they tried to translate alien symbols, the novel amorphous shapes used in the main experiment. These symbols could refer to either colors or sounds, equivalent to the category labels used in the main task, and could come from one of two similar alien languages (Lapian or Sindarin), corresponding to the

previous confidence judgments.

In the observation phase, the apps first detected which one of the two possible made-up languages the symbols belonged to and then proposed a category for the meaning, maintaining the structure of the main task. Participants were informed that the two languages used fairly similar symbols and, therefore, it was possible that the apps could be mistaken both about the language they belong to and their meaning.

At test, participants were first asked two questions about the apps' probability of correct categorization conditional on a given detected language. Specifically, participants were shown a new symbol from the same language task and were shown that the apps had detected the same language. They were then asked to make a forced choice between the two different possible meaning categories. This was repeated for the two languages. Agents differed in who was more likely to be correct across the two languages, so the expectation was that participants would choose the meaning offered by the app more likely to be correct for the respective language.

Following the conditional test, participants stated their general app preference for future translation tasks involving these languages in a 2AFC format and could explain their reasoning in an open-question format.

Following the conditional test, participants were presented with a general app preference 2AFC. Participants were asked which apps they would choose to use for similar future translation tasks involving these languages. Like in the main task, they were also asked if they wanted to explain their decision in an open-question format.

The perceived marginal probabilities of the apps being correct were reported on a 0-100 scale for both apps. Participants then reported on a 0-100 scale the frequency with which apps detected the two languages, and made a 2AFC decision on the more common language for each app.

Lastly, just as in the main experimental task, participants completed 12 trials of the translation task themselves.

The same counterbalancing and randomization as in the main experiment was applied throughout.

Experimental Manipulation

Main task

The accuracy of both observed agents was always 70%. A moderately high level of accuracy was chosen since it is unreasonable to expect good calibration with poor performance, and it leaves enough variability in accuracy to warrant monitoring the confidence. The profile of the agent pairs differed across seven between-participants conditions that manipulated: the differences in marginal probability of being confident and the informativeness of their confidence.

The agents' confidence could vary so that they were calibrated (as confident as s/he is accurate; 70% confident), over (90% confidence) or underconfident (50%). Further, the informativeness of confidence (see equation 4.1) was either matched, or varied.

Table 4.1 shows the agent pairings for the eight experimental conditions. We will refer to the agent who is more metacognitively attuned (confidence is more informative and/or better calibrated), who we predict agents will choose to collaborate with, as the reference agent.

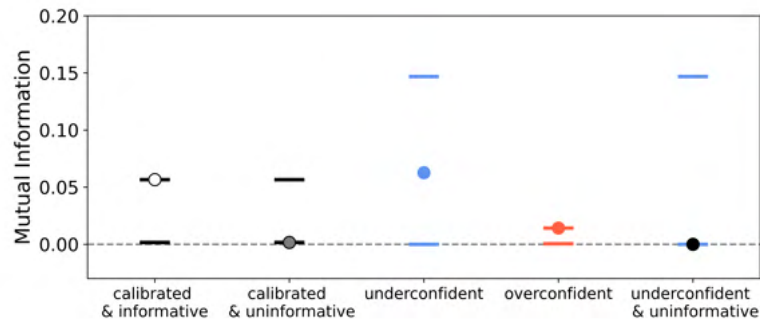
Across the first four conditions (A-D), the reference agent was confident 70% of the time and had maximally informative confidence given the constraint of behaving in a Bayes-rational manner (the probability of being correct given low confidence being at least 50% and lower than the probability of being correct given high confidence). The comparison agents were matched as closely as possible for informativeness in conditions A and B, but they were biased towards underconfidence (A) or overconfidence (B). Relative to the calibrated agent, the overconfident agent has a lower probability of being correct when confident, and the underconfident agent had a higher probability of being correct when confident.

On the opposite, the comparison agent in condition C was unbiased, but their confidence and accuracy were independent, their probability of being correct was the same for high and low confidence. In condition D, the agent was underconfident and maximally uninformative given their marginal confidence (so less informative than the underconfident agent in A).

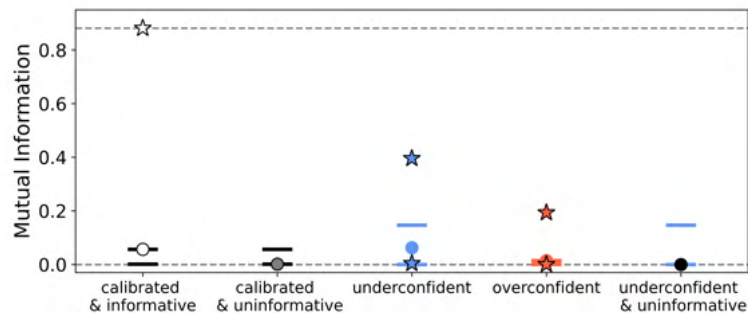
Table 4.2 and Table 4.3 crosstabulate the raw accuracy and confidence judgments for the agents used all conditions, alongside the associated probabilities, and type II hit and false alarm

Figure 4.2: Mutual information of agents.

(a) Conditions A-D of the experiment. The intervals marked with horizontal lines corresponding to the minimum and maximum MI that can be achieved with a given marginal proportion of being confident. The minimum and maximum have been computed only for plausible rational agents.



(b) Agents from conditions E-H (represented by star symbols) are added to the previous plot for ease of comparison.



(c) Difference in MI within agent pairs across conditions.

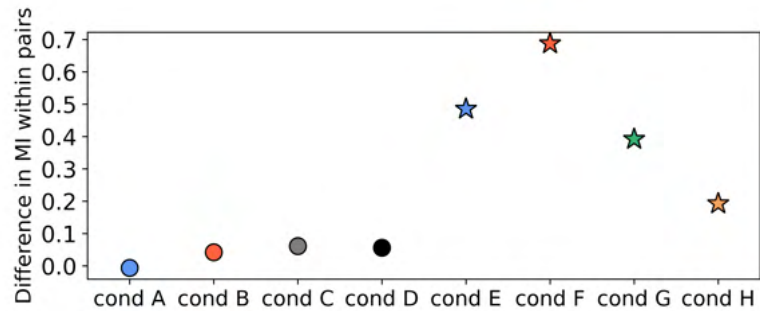


Table 4.1: The confidence and informativeness (mutual information, MI, in bits) of agents across the experimental conditions. The maximal information gain achievable is .88. Larger differences in MI are expected to result in more extreme preferences.

| Condition | Reference agent | | Comparison agent | | MI_{diff} |
|-----------|-----------------|-----|------------------|-----|-------------|
| | P(conf) | MI | P(conf) | MI | |
| A | .7 | .06 | .5 | .06 | .00 |
| B | .7 | .06 | .9 | .01 | .04 |
| C | .7 | .06 | .7 | .00 | .06 |
| D | .7 | .06 | .5 | .00 | .06 |
| E | .7 | .88 | .5 | .40 | .48 |
| F | .7 | .88 | .9 | .19 | .69 |
| G | .5 | .40 | .5 | .01 | .39 |
| H | .9 | .19 | .9 | .01 | .19 |

rates.

The rationality constraint greatly limits the differences in informativeness that can be achieved, as can be seen in Figure 4.2. To exaggerate differences in informativeness between agents, two additional conditions were constructed in which the reference agent was omniscient, that is, she was always correct following high confidence and always incorrect following low confidence. This agent was compared to an under (E) and overconfident (F) agent who had informative (but not as informative) confidence judgments. These two uncalibrated agents were further used as reference agents in two conditions in which they were contrasted with agents with the same bias, but poorer informativeness (G and H respectively).

Planned control conditions: Perception of relative accuracy

The control was performed on conditions E and F to test whether, if at all, the perception of accuracy is modified by over- and under confidence, respectively. These two conditions were chosen since they are likely to provide the highest chance of observing differences in perceived accuracy given that the agents differ in both marginal probability of being confident as well as the informativeness of confidence.

Post hoc control conditions: Perceptibly of statistical differences

Conditions C and F were chosen for the sanity check on whether the statistical differences between the two agents were large enough to be perceptible since we observed the largest and

smallest effect in preferences for these two conditions, respectively.

Materials

The shapes were generated as combinations of integral dimensions based on code provided by Op de Beeck et al. (2001) and are presented in Supplementary Information Figure A.4.18. The avatars are shown in the Supplementary Information Figure 4.19.

A pilot version of the experiment can be viewed on Pavlovia, the service used to host the experiment and data:

<https://run.pavlovia.org/Oana/categorization>. The data are downloadable from: tinyurl.com/58jzwhe8.

Table 4.2: Agent calibration profiles in Experiment 1 for conditions A-D

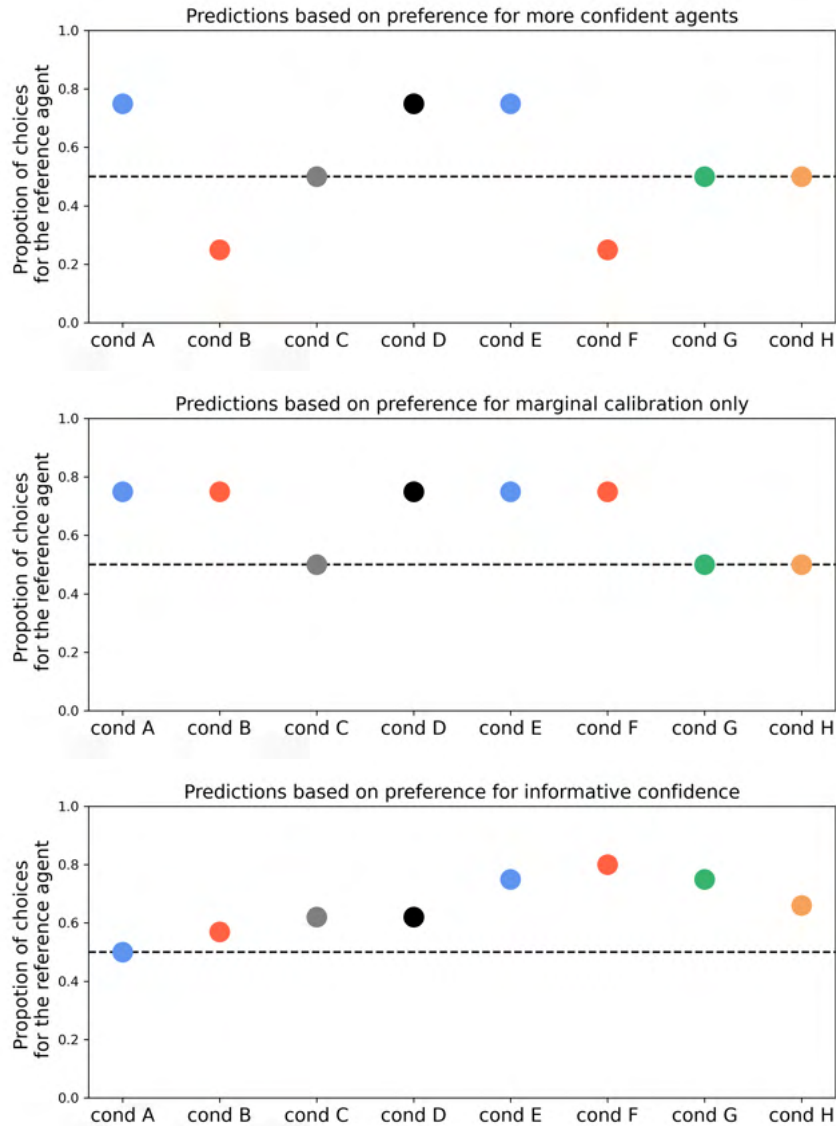
| Confidence | Realistically calibrated | | Calibrated Uninformative | | Under-confident | | Uninformative Confidence | | Over-confident | | Total |
|-------------------------------|--------------------------|-----|--------------------------|-----|-----------------|-----|--------------------------|-----|----------------|-----|-------|
| | high | low | high | low | high | low | high | low | high | low | |
| Decision accuracy | 33 | 9 | 29 | 12 | 25 | 17 | 21 | 21 | 39 | 3 | 42 |
| correct | | | | | | | | | | | |
| incorrect | 9 | 9 | 12 | 6 | 5 | 13 | 9 | 9 | 15 | 3 | 18 |
| Total | 42 | 18 | 42 | 18 | 30 | 30 | 30 | 30 | 54 | 6 | 60 |
| p(correct) | .7 | | .7 | | .7 | | .7 | | .7 | | |
| p(high confidence) | .7 | | .7 | | .5 | | .5 | | .9 | | |
| p(high confidence correct) | .79 | | .69 | | .59 | | .5 | | .93 | | |
| Type II hit rate | | | | | | | | | | | |
| p(high confidence incorrect) | .5 | | .7 | | .28 | | .5 | | .83 | | |
| Type II false alarm rate | | | | | | | | | | | |
| p(correct high confidence) | .79 | | .69 | | .83 | | .7 | | .72 | | |
| p(correct low confidence) | .5 | | .7 | | .57 | | .7 | | .5 | | |

Table 4.3: Agent calibration profiles for Experiment 1 for conditions E-H

| Confidence | | Perfectly Calibrated | | Underconfident | | Overconfident | | Total |
|---|--------------------------------|----------------------|-----|------------------------|--------------------------|------------------------|--------------------------|-------|
| | | | | | | | | |
| | | high | low | Informative confidence | Uninformative confidence | Informative confidence | Uninformative confidence | |
| Decision accuracy | correct | 42 | 0 | 30 | 12 | 22 | 20 | 42 |
| | incorrect | 0 | 18 | 0 | 18 | 8 | 10 | 18 |
| | Total | 42 | 18 | 30 | 30 | 30 | 30 | 60 |
| p(correct) p(high confidence) p(high confidence correct) (Type II hit rate) p(high confidence incorrect) (Type II false alarm rate) p(correct high confidence) p(correct low confidence) | p(correct) | .70 | | .70 | | .70 | | .70 |
| | p(high confidence) | .70 | | .50 | | .50 | | .90 |
| | p(high confidence correct) | 1 | | .71 | | .52 | | .81 |
| | p(high confidence incorrect) | 0 | | 0 | | .44 | | .89 |
| | p(correct high confidence) | 1 | | 1 | | .73 | | .70 |
| | p(correct low confidence) | 0 | | .4 | | .67 | | .67 |

4.3.3 Predictions

Figure 4.3: Visualization of predictions for collaborator preferences. Absolute values are only illustrative.



If the main driver of partner choices is the informativeness of conditions, the strength of the preference for the better calibrated agent should follow the within agent pair mutual information differences. Specifically, preferences for the reference agent should be observed across the board, including in conditions C, G, and H. However, if participants prefer agents who are more confident, then we should observe a preference for the reference agent in conditions A, D, and E; a preference for the comparison agent in conditions B and F, and no difference across

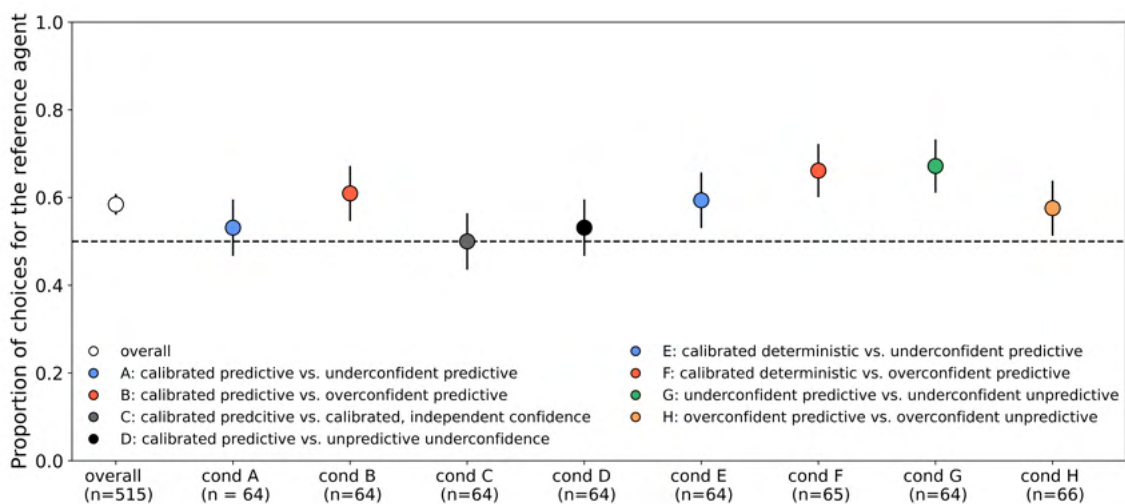
conditions G and H. Alternatively, if marginal confidence calibration is the basis of judgments, then we expect to find preferences for the calibrated reference agent in conditions A,B,D, E, and F but chance performance in conditions C, G and H. The different predictions are visualized in Figure 4.3. Aggregate preferences across these three groups will be tested against chance.

4.3.4 Results

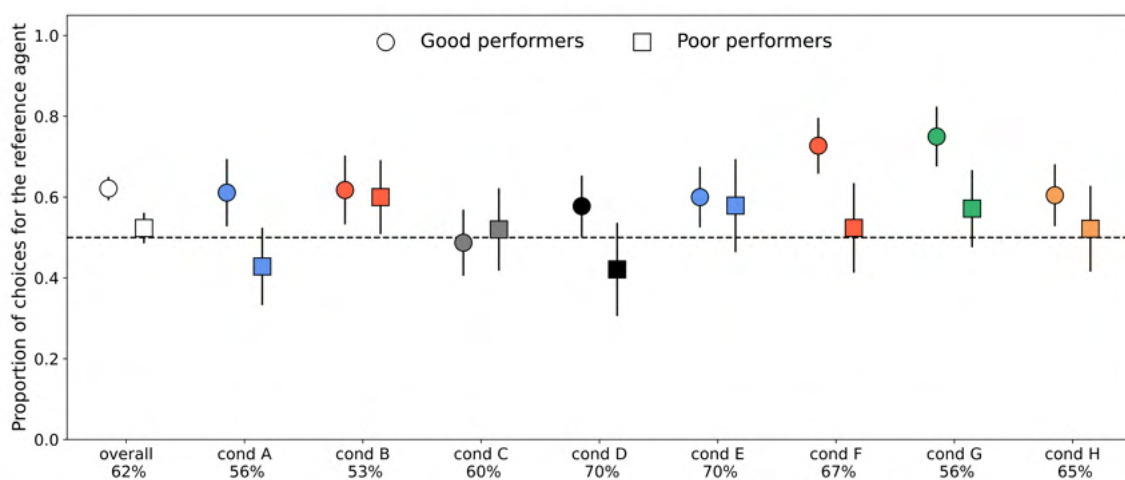
Main task

Figure 4.4: Collaborator choices across the conditions

(a) Proportion of times that the reference agent was preferred by participants. Lines denote \pm SE.



(b) Choices broken down by the participant's performance on the categorization task. Percentages under the conditions refer to the number of participants who scored above 75% correct (good performers) on the categorization task.



Overall, participants chose the agent who had better metacognition (referred to as the reference agent) 58.45% of the time, which was significantly above chance, $z(514) = 3.89, p = .001, BF_{alt} = 87.50$, Cohen's $h = .17$. As illustrated in Figure 4.4a, across all conditions where participants had a preference, the direction consistently favoured the reference agent as predicted. When the comparison agent was overconfident (conditions B and F), 63.57% of choices favored the calibrated reference agent, $z = 3.20, p = .001, BF_{alt} = 12.94$. The reference preference was also found when the biased agent was underconfident (conditions A, D, and E), although it was not significantly different from chance, 55.21%, $z = 1.45, p = .15, BF_{null} = 3.93$. Lastly, agent comparisons in which the calibration bias was matched and agents differed only in the informativeness of confidence (conditions C, G, H) led to a preference for the reference agent 58.25% of the time, $z = 2.33, p = .02, BF_{alt} = 1.28$. Continuous differences in the informativeness of paired advisers significantly predicted choices for the reference agent in a logistic regression, $\beta = .79(95\%CI : .046 - 1.53), p = .03$.

When adjusting the multiple comparisons for FWER ($\alpha = .015$), only the preference for the calibrated over the overconfident agents remained significant. However, a χ^2 test for homogeneity failed to find a difference between preferences across the three contrasts, $\chi^2(2) = 2.22, p = .33, BF_{null} = 22.94$, suggesting that preferences for the agent with better metacognition were equally strong across all types of comparisons, whether they involved an under/overconfidence bias or a loss of informativeness.

Moreover, preferences were more extreme for participants who were more attentive (or motivated) during the experiment. Figure 4.4b groups participants based on whether they scored 75% and higher¹⁰ on the final categorization task. Good performers chose the reference agent 62.12% of the time, and poor performers made the same choice only 52.08% of the time, $\chi^2(1) = 4.57, p = .03, BF_{alt} = 1.87$.

For good performers, the proportion of times the reference agent was chosen was above chance when compared to underconfident agents (conditions A, D, and E, $prop = .60, z(125) =$

¹⁰75% corresponds to the performance that is statistically different from chance on this task. All analyses were repeated including only participants who scored over 50% on this categorization task. The same qualitative pattern of results was found.

2.18, $p = .03$) and overconfident agents (conditions B and F, $prop = .68, z(77) = 3.40, p < .001$). The proportion of reference partner choices was also above chance when marginal probabilities were matched (conditions C, G, and H), $prop = .61, z(117) = 2.45, p = .01$. Lastly, when (relatively) matched for informativeness (conditions A and B), there was also a preference for a marginal match between accuracy and confidence, $prop = .61, z(69) = 1.96, p = .05$.

However, there are two notable departures from predictions based on informativeness in conditions C and D, where there was no consensus on preferences, $prop = .54, z(83) = .66, p = .51, BF_{null} = 5.97$. The results for condition D are extremely surprising given that this condition was predicted to elicit the largest effect as it pools together the effects of both marginal proportions of being confident and informativeness. Further, results contradict the preferences found in conditions G and H, where informativeness is the only explanation for the observed effect. We explored two explanations for this result: either the statistical manipulation was not sufficiently salient or preferences are modulated by another variable.

Visual exploration of the data indeed revealed considerable and unexpected differences in performance as a function of the interaction of the experimental condition and the gender of both the participant and the avatars. As illustrated in [Figure 4.6.1](#), the small aggregated preference effect was sometimes the result of averaging strong preferences (e.g., 87.50% of men preferred the calibrated agent in condition F). We did not expect this interaction and did not have specific a priori predictions for gender effects as a function of the condition or avatar, so we will not perform statistical analysis on gender effects, an analysis that would also be underpowered.

However, since these differences are sometimes large, and could be of interest for future research, we graphically present the data in the Supplementary Information [Figure 4.6.1](#). Overall, the strongest preferences for the reference agent were found for male participants judging male avatars (69.53% in favor of the reference agent) compared to male participants choosing between female avatars (55.12%) or female participants deciding between female (54.47%) and male (54.20%) avatars.

The average performance on the categorization task was 76.55% ($SD = 24.24, Median =$

83.33), slightly higher than that of the observed agents. The mean estimated accuracy of the chosen agent was 72.11% ($SD = 11.07$, $Median = 73.00$), a fairly close group estimate that showcases the wisdom of the crowd effect. Figure 4.5 illustrates pooled participant performance on the estimation and categorization tasks.

Importantly, no gender effects, referring to either the avatars' or participant's gender, were found in the categorization performance following the experimental task, nor in the estimated accuracy of the selected agent (in order, for participant gender: $t(512) = .59$, $p = .55$, $BF_{null} = 8.61$; $t(512) = -1.46$, $p = .14$, $BF_{null} = 3.62$; for avatar gender: $t(512) = 1.22$, $p = .22$, $BF_{null} = 4.95$; $t(512) = .53$, $p = .59$, $BF_{null} = 8.90$; see Figure 4.6).

Figure 4.5: Estimates of the agent's accuracy and performance on the categorization task.

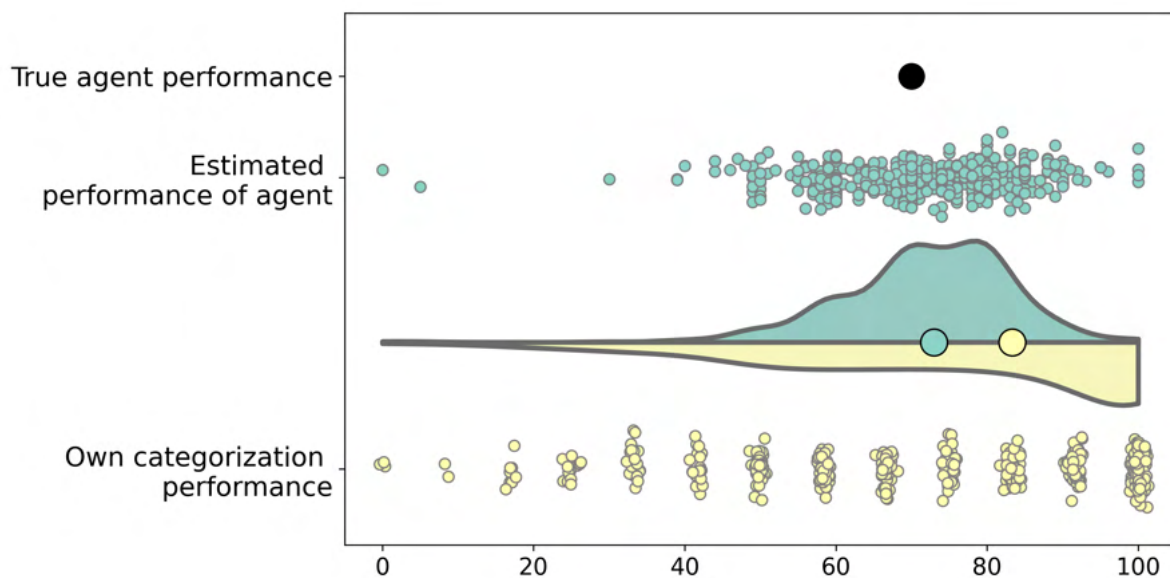
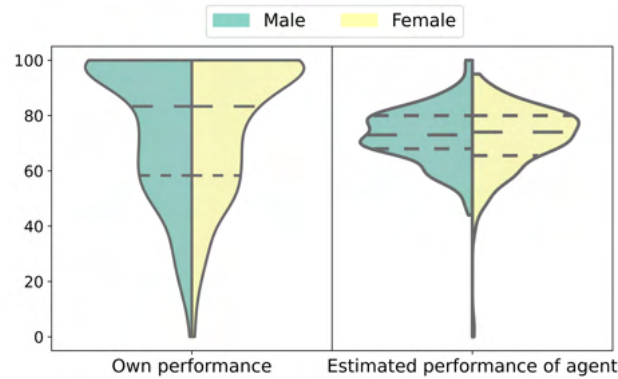


Figure 4.6: Violin plots showing the participant's estimates for the agent's accuracy and performance on the categorization task by participant gender.



Planned control task

Figure 4.7: Participants' estimates of the two agents' accuracy in the control conditions (replicating main task conditions E and F). Crosses denote sample means. In the rightmost plot of within participant differences against zero, circles denote means and lines correspond to medians.

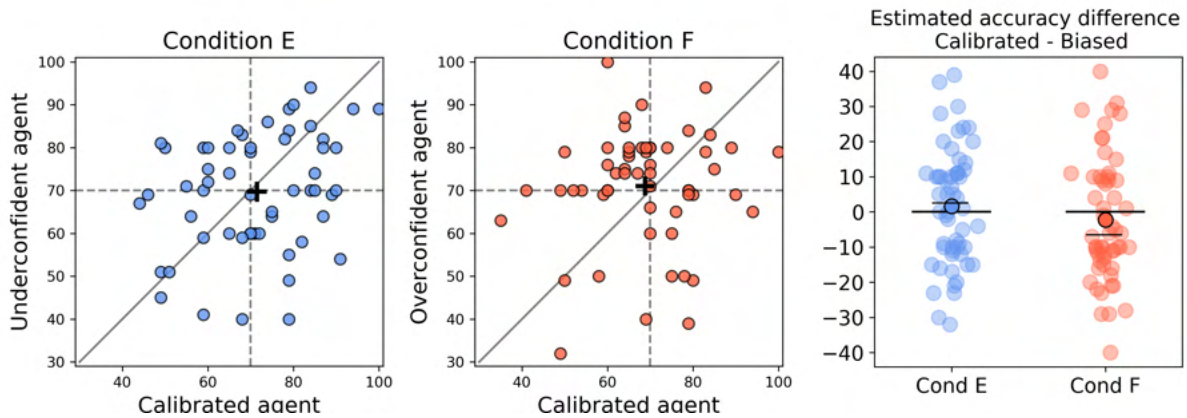


Figure 4.7 contrasts the participants' accuracy ratings for the reference calibrated agent and the biased agents in conditions E and F. Across the two conditions, estimates for both agents were centered around the ground truth value of 70%, although there was considerable variation in estimates. The average within-participant difference in the estimated accuracy for the two agents did not differ from zero in condition E, $M_{diff} = -2.28, SD = 17.39, t(55) = -.97, p = .33, BF_{null} = 4.35$, or F, $M_{diff} = 1.67, SD = 16.42, t(56) = .75, p = .45, BF_{null} = 5.24$. The same pattern of results was obtained using median tests.

There was no significant difference between the relative accuracy ratings of men and women in condition F, $t(54) = .70, p = .49, BF_{null} = 2.99$. However, a small, but significant difference was found in condition E, $t(53) = 2.05, p = .04, BF_{alt} = 1.50$, as women rated the calibrated (more confident) agent as slightly more accurate ($M_{diff} = 6.11, SD_{diff} = 15.16$), while men did not ($M_{diff} = -2.75, SD_{diff} = 16.45$), showing instead a small bias towards the underconfident agent. However, when looking only at participants who performed well in the categorization task, the bias observed in women disappeared. No significant differences were found as a function of the gender of the avatars.

In the preference question following the estimation task, participants were more likely to choose the agent whom they rated as more accurate, $prop = .81, BF_{alt} = 143.62$. Overall preferences within either condition were not different from chance, $BF_{null} = 4.46$. This result is to be expected given that on average the agents were perceived as being equally accurate, and assuming that participants expressed preferences that were consistent with whom they explicitly rated as more accurate. Thus, preferences for the more accurate agent would average to chance if participants are trying to be self consistent.

Post-hoc control task

Participants completing condition C were at chance in the two questions testing the comprehension of conditional probabilities: 26.56% of participants responded correctly to both questions relative to the 25% chance level for answering both questions, $BF_{null} = 7.00$. However, in condition C differences were observed between the two conditional questions. 64.06% of participants got the right answer for the question conditioning on the less common language, but only 45.31% of participants were correct for the remaining conditional, $\chi^2(1) = 3.81, p = .05, BF_{alt} = 3.05$. This finding is in line with the fact that the ground truth between-agent difference in the probability of being correct was larger within the rarely occurring language (.7 vs. .5) compared to the frequently occurring language (.79 vs. .69). If only this conditional difference was perceived, then preferences should favor the app perceived to improve in accuracy given the low (rather than high) frequency language. In condition C, this would result in a preference for the comparison app. This prediction was borne out by the data, as the participants

chose the less informative app in condition C 65.62% of the time, $z = 2.63, p < .001, BF_{alt} = 3.53$. Moreover, participants who answered the conditional question correctly were more likely to choose the comparison app (78.05% vs 43.48%), $\chi^2(1) = 6.35, p = .01, BF_{alt} = 17.68$.

In condition F, participants performed above chance in the conditional probability test, .56, $BF_{alt} = 5.49e4$, and preferred the reference agent 59.37% of the time.

In the explicit ratings of the agent's accuracy, there was no significant difference between the estimated accuracy of the two agents, but, as before, the variability within the sample was fairly large. In condition F, the mean difference was -.10 ($SD = 16.97, Median = 0$), $t(47^{11}) = -.04, p = .97, BF_{null} = 4.99$, and similar in condition C the observed difference was -3.77 in favor of the reference (calibrated) agent ($SD = 21.16, Median = -4.50$), $t(63) = -1.42, p = 1.60, BF_{null} = 2.81$.

For both conditions, most participants correctly reported that the apps detected one language more often than the other (corresponding to the confidence marginals in the main task), 71.87% of the time in condition F, and 67.18% in condition C, overall $z = 4.80, p < .001, BF_{alt} = 2303.39$. Furthermore, in estimating the frequency with which apps detected the languages, we found no differences for condition C in which the probabilities were equal by design, $M_{diff} = -4.66, SD_{diff} = 26.77, t(63) = -1.38, p = .17, BF_{null} = 2.96$. On the contrary, there was a significant difference in condition F in line with the fact that one language was detected 20% more by one app than the other, $M_{diff} = -7.67, SD_{diff} = 19.84, t(47) = -2.65, p = .01, BF_{alt} = 3.54$. Lastly, the agent preferred was perceived to be more accurate, but not more confident.

4.3.5 Discussion

Descriptively and qualitatively, the pattern of results generally tracks predictions based on the informativeness of the participants (see [Figure 4.3](#)). Moreover, participants' preferences for future collaborators were just as extreme in conditions in which the two agents were matched

¹¹Unfortunately, we could not record the continuous measures for accuracy and marginal language detection for the first 12 participants due to a technical error.

for bias, suggesting that choices were modulated by differences in the informativeness value of the confidence judgments. This could only be possible if participants were monitoring not only the marginal probability that agents were correct or confident, but also the conditional relationship between accuracy and confidence. However, effect sizes proved to be rather small, suggesting that tracking informativeness across time is a difficult task, and one that people could routinely fail at inside and outside the lab. On the other hand, participants in the current study had no reason to believe that the potential informants had any interest in misleading them given the collaborative cover story, so we expect the metacognitive abilities of informants will be scrutinized more in settings where there is a potential for intentional manipulation of confidence.

Contrary to the previous findings of Price and Stone (2004), and in accordance with the proposals of Sah et al. (2013) and Tenney et al. (2011), we found that the participants were consistent in their vigilance to expressions of overconfidence. There are several factors that could have made it more likely to find this effect under the current design: higher ease of tracking the covariance of binary expressions of confidence and accuracy across trials, larger differences in the biases of agents, and perhaps the different cover story.

Since the preference for a better calibrated agent was present both for comparisons with an underconfident and an overconfident agent, it does not reflect a preference for a more extreme, more confident agent as suggested by previous studies. Second, the preference is stronger for comparisons with the overconfident agent (prop = .63 vs. .51), which was predicted based on the informativeness of confidence, since the overconfident agent had confidence judgments that were less informative than those of the underconfident agent. Further, within the marginally matched conditions, the preference for the calibrated agent survived and was stronger for the comparison of the two underconfident agents (prop = .63) than for the comparison of the two overconfident agents (prop = .56). This is again consistent with predictions on the basis of confidence informativeness differences.

At the same time, when the contrasted potential advisers were equally informative, the one who was confident to the extent that she was correct was preferred, suggesting that marginal

probabilities of being confident are tracked as well. However, it is clear that there is not just a blanket preference for more confident advisors, as the more calibrated agent was preferred to an agent that was more likely to make highly confident judgments.

Furthermore, the preference for calibrated social partners was not accompanied by a group-level difference in their perceived accuracy. However, we did find a self-consistency/halo effect in the planned control task as participants preferred the agent whom they had previously rated as more accurate. This led the aggregate preference judgments in the control condition to be indistinguishable from chance, unlike in the main task. Based on this finding, it is likely that the preference judgments in the main task were not mediated by a bias in the perception of the accuracy of the observed agents.

It is true, however, that the control conditions comparing accuracy were not within-participant, so we can only draw conclusions about group-level differences. However, as pointed out before, there are issues with a within-participant measurement as well, as they can lead to biasing due to post-hoc justification. It seems very unlikely that small error in perceived accuracy could then lead to preferences for calibrated agents that are this consistent across multiple manipulations.

There was a surprising result as the more metacognitively sensitive agent was not preferred when contrasted with an agent with confidence judgements independent of accuracy. A replication of the experiment with a nonsocial cover story offered a likely explanation. The findings indicated that the participants did not detect one of the conditional differences. Their preferences were, however, fully consistent only with the partial evidence they obtained.

The effects of the gender of the participant and the avatar were surprising and warrant further targeted investigation. Price and Stone (2004) found no differences between pairs of male and female financial advisors in one experiment, but no analyses were presented that included both the gender of the avatar and the participant's. They also found a small difference as a function of the gender of the participants, as more men preferred financial advisors with more extreme confidence. This is partially consistent with current results, but does not explain the gender effects across all conditions. The gender effects may partially explain the fairly small

effect size observed. The fact that the current experiment found an effect of avatar gender is surprising given that the avatars used here are highly schematic and their gender was not made salient.

To conclude, confidence was shown to be a useful signal that informs partner choice for cooperation tasks. However, the results need to be replicated and extended with other tasks in order to argue that tracking of calibrations is implicit and spontaneous. Although participants were not told to monitor the covariation of confidence and accuracy, they were also not engaged in a secondary task and had little else to focus on in the display, which made it obvious that the purpose of the observation phase was to remember the confidence and accuracy judgments of the participants. On the other hand, in such a simple scenario, analysis of the open ended questions revealed that even for participants who justified their partner selections with reference to extraneous facts (e.g. ‘I liked the person better’, ‘I like the color of their shirt’), the probability of choosing the calibrated agent was above chance. It is intriguing whether monitoring of confidence informativeness is an automatic and implicit process similar to statistical learning, or whether it only occurs in a collaborative (or pedagogical) context or, generally, like here, in a setting where several individuals are being contrasted.

4.4 Experiment 2

4.4.1 Introduction

Experiment 1 showed that participants can determine the relative metacognitive ability of two potential collaborators in a categorization task, by tracking across time the statistical relationship between their accuracy and confidence. In the current experiment, we test whether, in a similar task, inferences about basic biases of confidence expressions (under and over estimation relative to accuracy) lead not just to general collaborator preferences, but also extend to differences in the willingness of accept the suggestions of competing advisers.

A substantive test of metacognitive monitoring should involve checking the extent to which humans are making optimal adviser choices in light of their metacognitive abilities. Specif-

ically, whether people leverage the functional mapping between accuracy and confidence in order to determine, on every given encounter, the probability of advisers being correct given their stated confidence (and any other potential contextual information). We refer to this quantity as recalibrated confidence, in contrast to the explicitly stated confidence of advisers. The explicit and recalibrated confidence are the same only for calibrated advisers.

We propose a simple experiment in the judge-adviser framework in which participants could infer the relationship between the accuracy and confidence of two agents (manipulated across conditions) by observing them repeatedly perform a novel task. Following this, participants made multiple decisions relying solely on disagreeing advice from the potential advisers. We hypothesized that trial-by-trial, participants will choose the suggestion of the adviser with the highest recalibrated confidence as opposed to the highest stated confidence. In the current experiment, the optimal recalibration strategy leads to sometimes selecting the advice of an adviser who is explicitly less confident than their competitor independent of calibration. Thus, current predictions sometimes disagree with both the confidence heuristic and the calibration hypothesis.

A simple example can be used to motivate the experimental task. Imagine you are a student struggling with solving an equation. You have two friends, Anna and Emma, who both scored 70% in the latest math test. Anna says she is 90% confident she can solve it, while Emma rates her chances at 70%. They did equally well on the test, so who do you ask for help? At face value, given an assumption that accuracy and confidence go hand in hand, you should ask Anna to help. However, if you also know that after the last math test Anna thought she was 100% correct and Emma's confidence was at 60%, you can factor in Anna's overestimation and Emma's slight underestimation. Chances of Anna getting it right are likely around 60% and Emma's around 80% so in the end you are better off asking Emma for help.

Participants made such decisions in the experiment, comparing a calibrated agent to either an underconfident or overconfident, but equally accurate, agent in between participants conditions. Here, over and underconfidence are biases of estimation (vs. precision or relative placement). Further, in within-participant control conditions, participants compared two ad-

visers who had the same accuracy and confidence, and as well two agents who had the same confidence, but differed in accuracy. These comparisons functioned as controls to establish the internal validity of the task. Potential advisers only differed by a constant bias, as the sensitivity of linear relationship between accuracy and confidence was kept constant, making it as easy as possible to apply the recalibration.

A pilot study was conducted and is briefly presented in Supplementary Information [subsection 4.6.3](#).

4.4.2 Methods

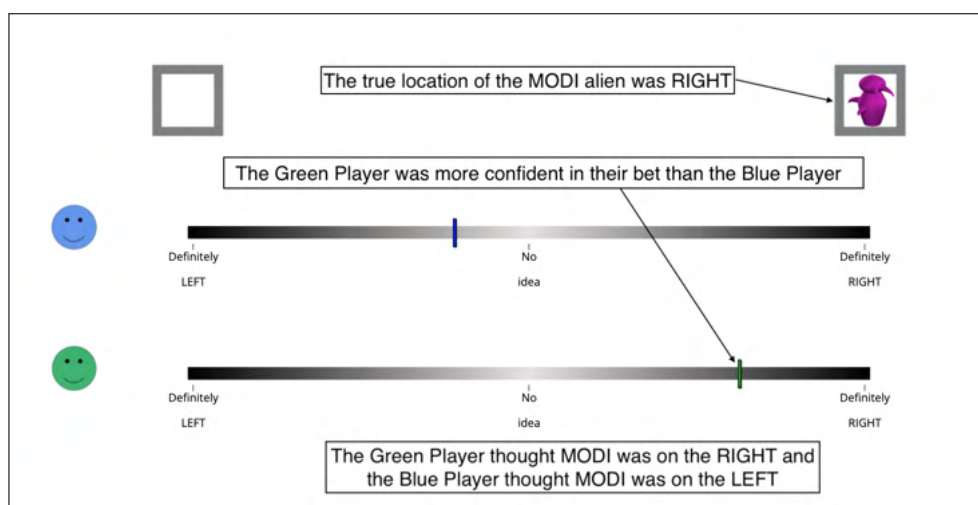
Participants

60 participants (29 female, 1 nonbinary, 30 male) were recruited online through the Prolific online platform. Participants were evenly split across two between-participants conditions. The mean age of participants was 33.27 years ($SD = 11.31$, range: 18 - 66 years old). English was the first language of all participants and the majority resided in the UK.

Participation was rewarded with £6.5, including a participation reward of £5.5 and a performance-based bonus of £1.

Ethical approval was obtained from the United Ethical Review Committee for Research in Psychology (EPKEB) in Hungary.

Figure 4.8: Example trial from the observation phase.



Task

At the start of the experiment, participants observed a pair of agents performing a simple novel categorization task across several trials. Participants then engaged in a betting task in which they had to make decisions based on the advice of these agents. Lastly, participants answered questions about the performance of the two observed agents and stated their overall preference for one of them.

During the **observation phase**, participants saw two other fictitious participants get tested on a binary categorization task while stating their confidence in the correctness of their response on a continuous scale. Specifically, the agents were betting one virtual coin every trial on whether a “Modi alien” was depicted in the image on the left or on the right side of the screen. It was stressed that these two agents were not communicating and could not see each other, but were performing the task individually on the Prolific platform. .

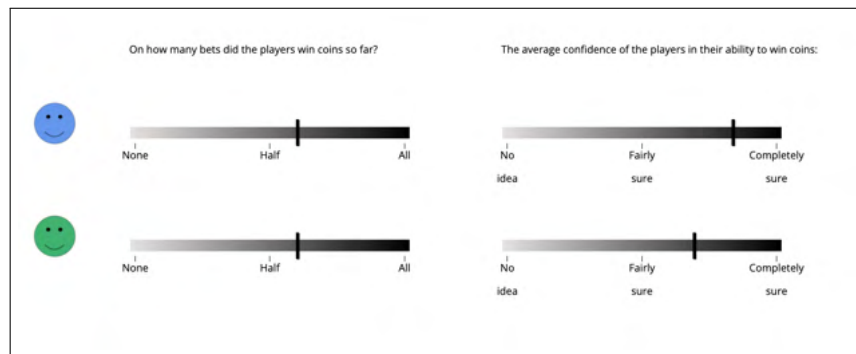
The agents expressed their answer and confidence simultaneously: the direction of the slider position relative to the center of the scale marked their decision (alien is left or right) and the distance from the center of the slider marked how confident the participant was (see Figure 1). Only verbal labels were presented on the scale such that the center was “I don’t know” and the extremes were marked with “Definitely right!” or “Definitely left!”. Further, a grayscale gradient was used to mark the increase in confidence. Following the agents’ decisions, the true location of the alien was presented on the screen. [Figure 4.8](#) illustrates a training trial.

The observation phases consisted of 120 trials. Randomly interleaved attention checks were presented following 10% of trials. Participants were asked to make a 2AFC choice about whether a given agent was correct in their answer. Feedback was offered.

After every 30 trials, participants were shown the number of correct answers and the average confidence for the two agents thus far. Both accuracy and confidence were presented on verbally labelled continuous scales (see [Figure 4.9](#)). This means that participants were reminded of the summary statistics at the end of the observation phase.

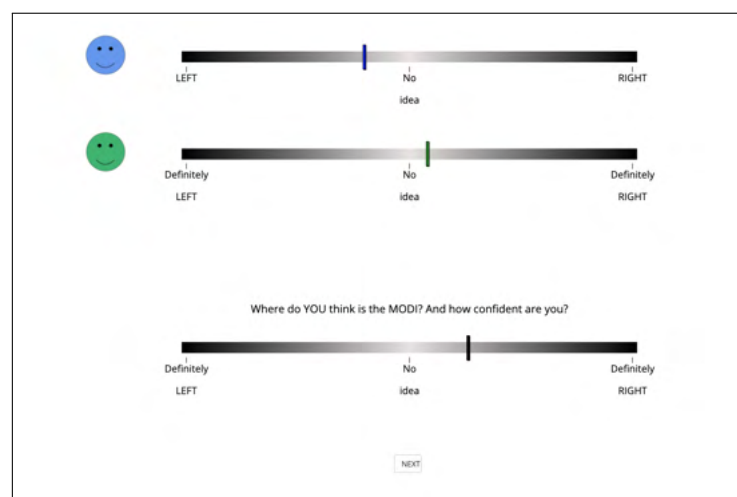
Decisions were presented alone on the screen for 2,500 ms, the true location was then added

Figure 4.9: Slide summarizing the performance of the two agents so far. These summaries were presented to participants every 30 trials.



and presented on screen for 2,000 ms and the intertrial interval was 1,500 ms. The answer to the check was not speeded, and the feedback was presented for 500ms. Participants were allowed to consult the summary boards for as long as they wanted.

Figure 4.10: Example trial from the Betting phase. Here, the participant chose the advice of the agent represented by the green avatar.











In the **betting phase**, participants were instructed that it was their turn to perform the categorization task. To incentivize performance, participants were told that on each trial they would have to bet one virtual coin on their answer, and as a function of that number of coins they won by the end of the task, they could receive a monetary bonus. Participants were only informed how many coins they earned at the end of the experiment. In total, participants made 60 bets.

As before, participants could see the decisions and associated confidence of the two agents from the observation phase. Importantly, participants were not shown the two images in which

the “Modi alien” could be depicted so they had no way of knowing the location of the alien by themselves. Thus, to make a ‘blind’ forced decision, participants could only rely on the decision made by the two agents and their associated decision confidence. Participants responses were made on a scale identical to the one used by the two agents (see [Figure 4.10](#)). As such, participants simultaneously made a decision and expressed their level of confidence in that decision. Answers were unspeeded and no feedback was provided.

In the **post-test phase**, participants were asked to estimate the accuracy and confidence of the two agents on continuous scale. Participants then made a 2AFC decision on their partner preference for a future similar collaborative task.

Figure 4.11: Task design. Test and Control blocks were run within-participant. Conditions A and B were between participant conditions. Numerical differences presented here were scaled for visual presentation in the experiment (50% confidence corresponding to the center of the scale, and 100% to the extremes).

| | TEST | | CONTROL | |
|--------------------|---|---|--|---|
| Condition A |  |  |  |  |
| Mean Confidence | 80 | 65 | 80 | 80 |
| Mean Accuracy | 80 | 80 | 80 | 80 |
| | calibrated | underconfident | identical & calibrated | |
| Condition B |  |  |  |  |
| Mean Confidence | 70 | 85 | 70 | 70 |
| Mean Accuracy | 70 | 70 | 70 | 85 |
| | calibrated | overconfident | accuracy difference | |

Design

Each participant completed the task described above twice, with two different pairs of (fictional) advisers, the order of which was counterbalanced. In addition to this within-participant manipulation, there was also a between participant manipulation of the agent pairings. [Figure 4.11](#) illustrates the design of the experiment.

We refer to one of the within-participant tasks as the Test block and one as the Control

block. Control blocks were used as sanity checks of the design, and provided a measure of the minimum and maximum effects that can be expected. The control block of condition A used two agents matched for accuracy and confidence. The same agent was presented twice, with shuffled trial order. Any differences observed in this block can only be attributed to perceptual noise. In the control block of condition B, the two agents had the same confidence, but one of them was more accurate. This modulation should induce strong preferences, so a failure to elicit a statistically significant effect would mean that participants did not learn the statistics of the task.

In the test block of condition A, participants observed an underconfident agent paired with a calibrated one. In the test block of condition B, an overconfident agent was displayed alongside a calibrated one. Importantly, in both test blocks, the two agents had the same overall accuracy and the same approximately linear relationship between accuracy and confidence (see [Figure 4.12](#)).

The confidence values of the agents in betting phase trials were selected to make it possible to distinguish whether participants chose trial-by-trial the agent with the highest recalibrated or explicit confidence. First, in all betting trials, the two agents disagreed about the location of the alien. Participants were explicitly informed that only trials with disagreements will be shown, since their decision would be obvious when the two agents agree¹². Second, we ensured that there were sufficient numbers of trials (33%) in which decisions based on explicit confidence differences were different than decisions based on recalibrated confidence, but shy of making decisions based on recalibration overwhelmingly favour one agent. Lastly, the average confidence of the agents was roughly equal in the betting trials.

The choices made by the agents throughout (left or right) were also randomized, with the constraint that an equal number of 'Left' and 'Right' decisions were made to avoid participants developing a location bias. The four fictitious agents were represented by abstract avatars that only differed in color, the assignment of which was counterbalanced across participants.

¹²Two well-performing agents, such as the ones presented in the experiment, by chance will have a high overlap of decisions, so the lack of agreement needed to be explained

Procedure

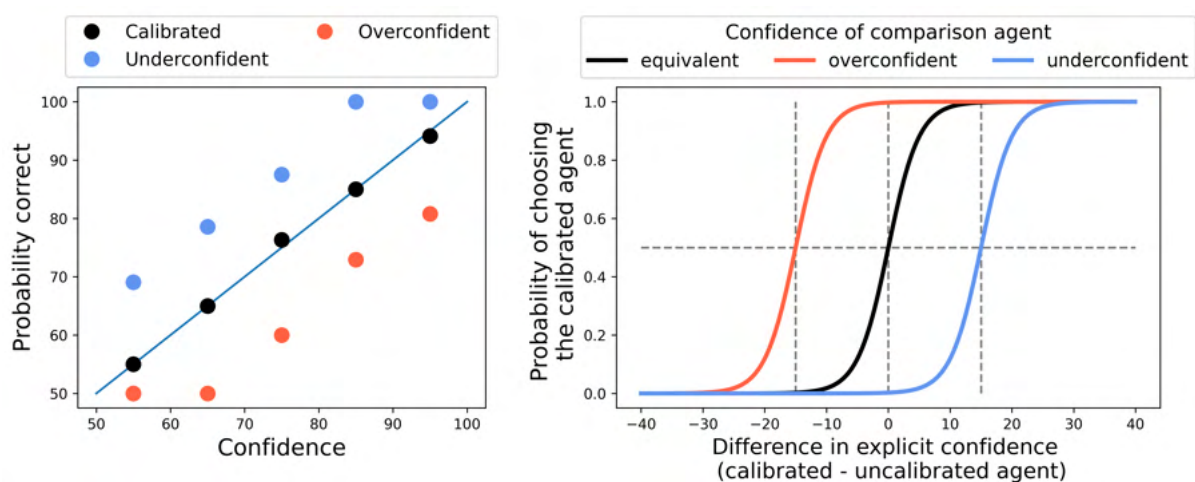
The experiment took on average 55 minutes (maximum 132 minutes), including a 5-10 minute break half way through the experiment. The break was used to decrease the likelihood that participants would carry over inferences about avatars previously seen at a given position over to the new avatars at the same position. As such, a minimum 3 minute break was enforced, but participants were allowed to take as long as 30 minutes.

Materials

The alien images presented during the betting phase were taken from different families of the symmetric Greebles dataset designed by Scott Yu. Images are presented courtesy of Michael J. Tarr, Carnegie Mellon University, <http://www.tarrlab.org/>

The experiment can be viewed here: <https://run.pavlovvia.org/Oana/categorization2> and the data are available at tinyurl.com/58jzwhe8.

Figure 4.12: Left: Relationship between confidence and probability of being correct for agents with the same accuracy, but different marginal confidence. Right: Predictions for decisions as a function of confidence differences between advisers and condition.



Predictions

The main prediction concerned the binary decisions of participants, namely how their choices should vary as a function of the confidence difference between the two advisers' confidence. The null hypothesis was that agents rely solely on explicit differences in confidence when choosing advisers, specifically, on a trial-by-trial basis choosing the answer proposed by the adviser who had a higher level of stated confidence on that trial. Alternatively, agents could make decisions based on differences in the agents' probability of being correct given their stated confidence (recalibrated confidence). For the conditions in which agents had equal accuracy, this results in the predictions in Figure 4. An underconfident agent is more likely to be accurate in our design (and therefore should be chosen) even when they are slightly less confident ($>15\%$) than the calibrated agent (Condition A). To the contrary, an overconfident agent should not be chosen when they are slightly more confident ($<15\%$) than a calibrated agent (Condition B). [Figure 4.12](#) illustrates the predictions for the participants' decisions based on advice.

Further, continuous confidence judgments were expected to mirror effects observed for the binary choices, assuming that participants compute their own confidence in each of the two possible answers and choose the one with the highest confidence.

Given that the estimation task and the overall preference test were presented following the betting phase, high consistency is expected, and as such it is possible that preference judgments would depart from those observed in Experiment 1. The intuition is that if participants, following our prediction, more often choose the advice of an underconfident agent, this may also bias them to prefer them. However, rationally speaking, participants should continue to prefer the calibrated agent for a future task, given that the ground truth accuracy might not be known in this task.

Data analysis

Participants were removed from the analysis based on two preset criteria. First, participants kept in the analysis had to respond above chance in the memory test. Second, for the main

analysis, only participants whose decisions varied with the confidence of the two advisers were included (there was strong consistency between exclusions based on these two criteria). In order to determine this, a logistic regression was fitted for every participant's bets using the bets and confidence of the two agents as predictors. AIC was used to compare this model to a random response model. While participants better fitted by the random model were not included in the analysis, we report their choices descriptively, as, for instance, it is potentially meaningful if they chose the same adviser consistently and ignored trial-by-trial variability.

A Bayesian generalized mixed effects logistic model was fit to the choice data using the pyMC3 package, (Salvatier et al., 2016), using the agent confidence difference as a fixed predictor and participants as a random intercept. The intercept and slope of each participant were assumed to be sampled from Gaussian distributions with unknown mean and standard deviation. The hyperpriors on the population mean and standard deviation were generic weakly informative priors following Gelman et al. (2003). The means were assumed to come from a Student's t distribution and standard deviations were sampled from the Half-normal distribution. Three chains with dispersed initialization were used for estimation, with 10^4 iterations each. Gelman-Rubin and Geweke diagnostic tests were used to check for convergence of the model. Posterior predictive checks were used to assess goodness of fit.

A model was fit separately for every condition, resulting in a posterior distribution for the categorization boundary.

As a sanity check that participant choices were indeed driven by the difference in the confidence of the two agents, and were not dominated just by just one of the agent's choices, the same logistic model was fitted using both agent's judgments as predictors. The relative weights assigned to the two agents were compared for statistical differences. Participants' decisions varied as a function of the difference in the confidence judgments of the two agents, as the weights used for the two agents were equal (see HDIs for population parameters, and individual weights in Figure 4.24), therefore, we present boundaries calculated based on regressing decision to explicit confidence differences.

Visual exploration of the data revealed that continuous (absolute) confidence judgments

were in fact mostly driven by the confidence of the chosen agent. A priori the expectation is that the confidence of the participant should be influenced by both the confidence of the agent whose advice was taken as well as by the confidence of the other adviser. Intuitively, the higher the confidence of the adviser whose answer was chosen, the higher the confidence of the participant in her answer. Conversely, due to the inherent disagreement in the adviser recommendations, the higher the confidence of the not chosen agent (in the opposite answer), the less confident should the participant be in their selected answer.

To explore the influence of the confidence of the chosen adviser and that of the other adviser on the participants' continuous confidence ratings, we regressed participants' confidence ratings on those of the agents and statistically compared the resulting weights, and tested whether the addition of both advisers' suggestions improved the variance explained.

Lastly, preference judgments and estimation of accuracy and confidence across the conditions were compared to chance and the ground truth, respectively.

4.4.3 Results

Eight participants were excluded from the experiment based on the low attention check performance and random response pattern (the two generally went hand in hand). A further 9 participants had only one condition included in the final analysis. Supplementary [Figure 4.22](#) presents the AIC comparison of the random and informed response model. For the purpose of visualizing individual variability in behavior, we present categorization boundaries from logistic regression fits for every participant. For each participant, a logistic regression was fitted separately on their decisions in the Test and Control betting trials with the explicit confidence difference between the agents as the predictor. This was equivalent to using the difference in the probability of being correct since the relationship between confidence and accuracy was approximately linear for the range of values used.

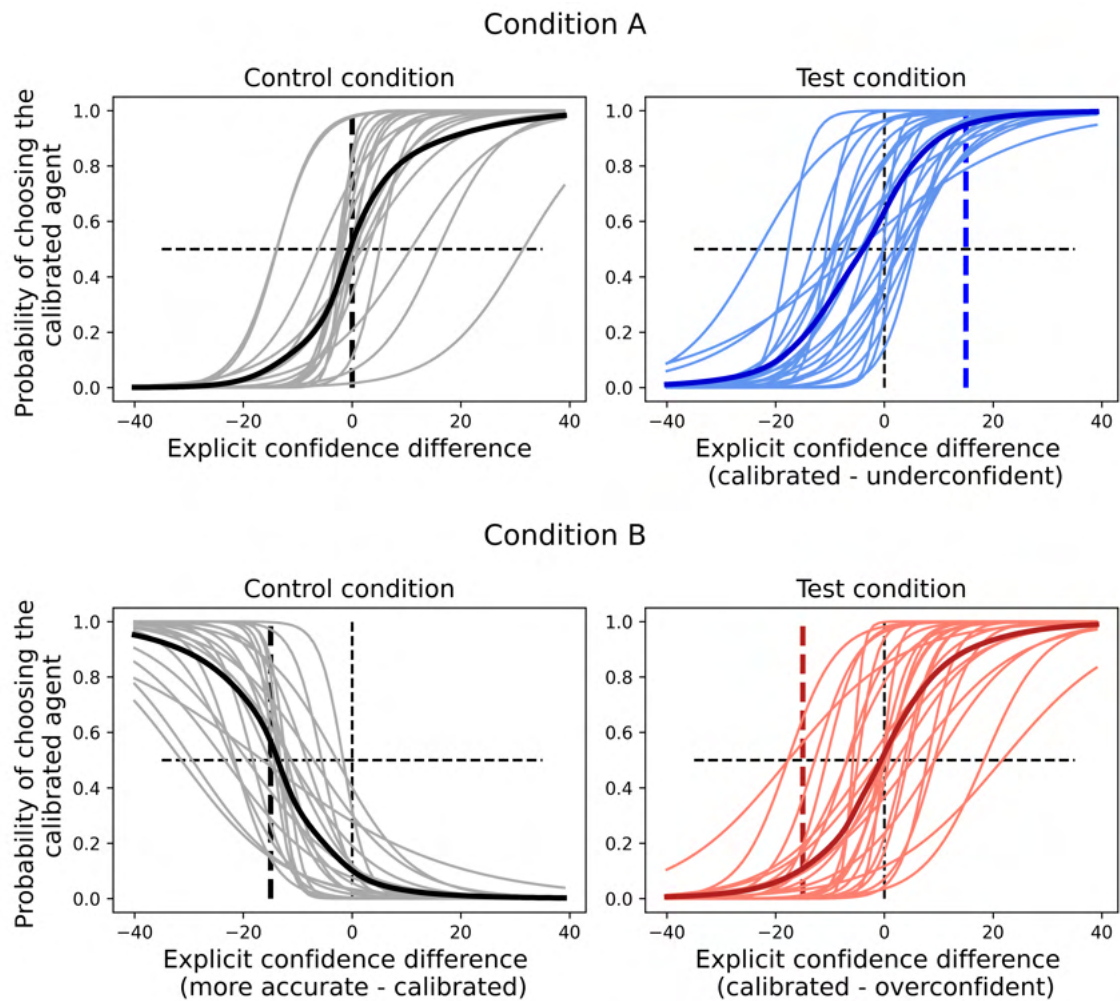


Figure 4.13: Fitted curves from each participant's logistic model on decisions. Dashed vertical lines represent predictions for the boundary based on recalibration and bold lines are group averages.

Adviser choices: Control Blocks

The most likely boundary for the control block of condition A, when the two presented agents were in fact identical, was -0.08 , 95% Highest density interval (HDI): $[-2.82, 2.69]$, see [Figure 4.14](#). As seen in [Figure 4.13](#), there was very little variability in the individual boundaries of participants, which closely clustered around zero, $M_{boundary} = 1.34, t(18) = .58, p = .57, BF_{null} = 3.62$.

In the control pair of condition B, when the two agents had the same confidence, but differed in accuracy, participants' boundaries shifted, as the more accurate agent was chosen be-

yond what would be expected from explicit confidence differences. The maximum a posteriori estimate (MAP) for the boundaries was -13.16, 95% HDI [-17.38,-9.55]. The predicted difference (15) was included in the HDI, but zero was not. Individual boundaries showed high consistency, $M_{boundary} = -14.73, t(21) = -8.97, p < .001, BF_{alt} > 10^5$.

Adviser choices: Test Blocks

Results did not follow predictions in the test blocks. In condition A, the boundary MAP estimate was negative, -3.77, 95% HDI [-6.83, -.85], suggesting that the more confident (and calibrated) agent's advice was used even when they were somewhat less confident than the underconfident agent. Individual boundaries were also predominantly negative, $M_{boundary} = -4.48, t(23) = -2.92, p < .01, BF_{alt} = 6.09$.

There was no discernible pattern at the group level in condition B, as individual boundaries varied widely around zero $M_{boundary} = -.16, t(25) = -.09, p = .93, BF_{null} = 4.81$. The MAP estimate for the boundary was -.43, 95% HDI [-3.22, 2.29].

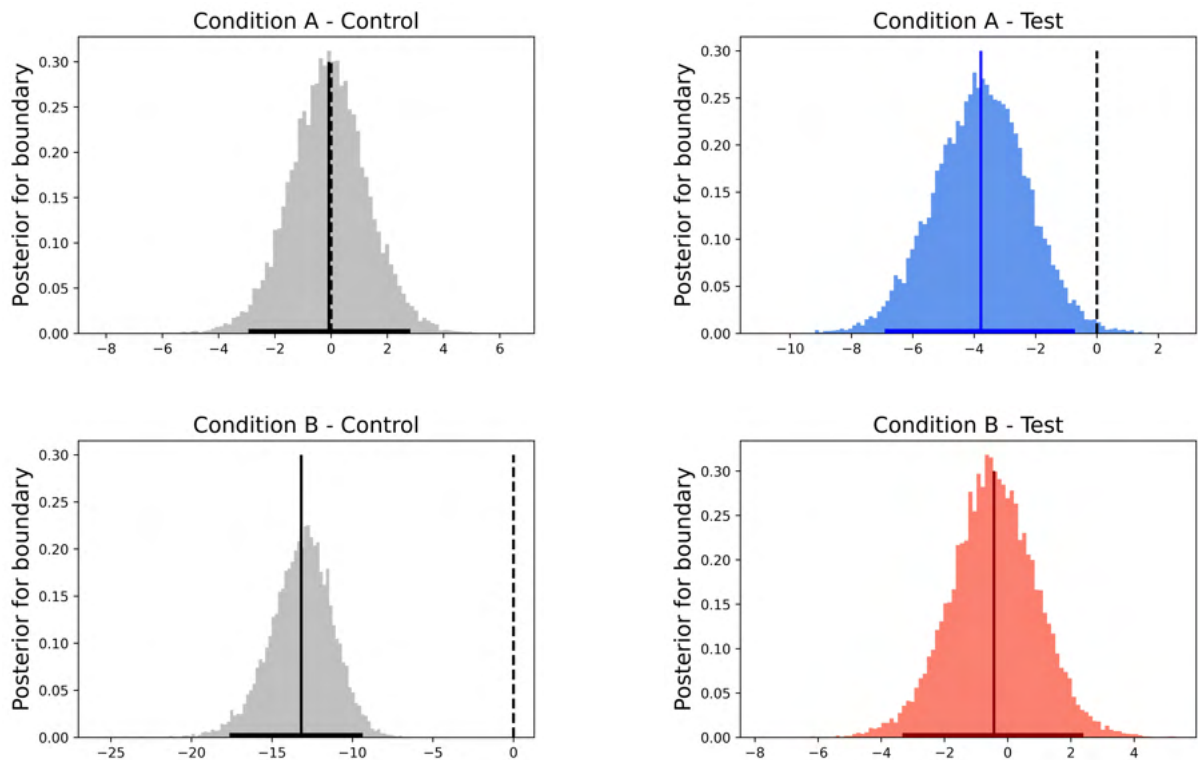


Figure 4.14: Posteriors distribution for the categorization boundary. Vertical lines represent the MAP and horizontal intervals include the 95% HDI.

Continuous confidence judgments

Contrary to predictions, the continuous confidence judgments produced by participants overwhelmingly depended on the confidence of the agent whose advice was taken, and the confidence of the other agent was not meaningfully incorporated in their stated confidence. This is evident from the fact that adding the confidence of the agent whose advice was not taken to the model did not increase the amount of explained variance for the vast majority of participants (Figure 4.15 and Figure 4.15 broken down by condition).

Figure 4.16 shows the relationship between the advice of the two agents and the continuous confidence judgments of the participants. Figure 4.23 shows the fitted weights for every participant, computed by regressing their absolute confidence on the absolute confidence of the two agents. In line with the R^2 values, the weights for the agent whose advice was disregarded were near zero.

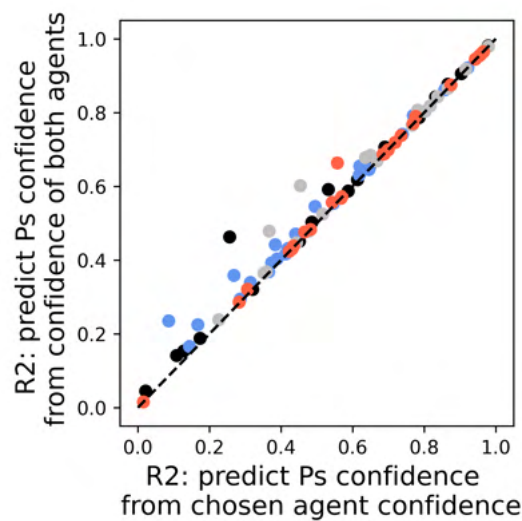


Figure 4.15: R^2 for predicting individual confidence judgments from the confidence of the potential advisers. Each dot is a participant, all conditions are overlaid (red: cond A test; blue: cond B test; gray: cond A control; black: cond B control).

Future collaborator preferences

When the two agents were identical, preferences for a future collaborator were at chance, $prop = .47, z = -.36, p = .71, BF_{null} = 2.30$. Participants preferred to collaborate with the

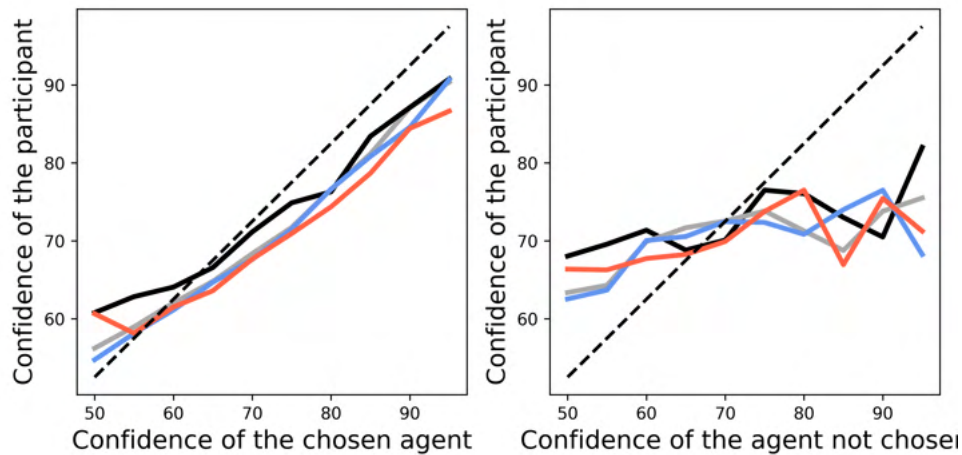


Figure 4.16: Relationship between confidence of chosen adviser and average confidence of participants.

more accurate agent of equal confidence, $prop = .90, z = 7.30, p < .001, BF_{alt} = 3279.49$.

In the test block of condition A, when deciding between an underconfident agent and a calibrated one with equal accuracy, the majority of the participants chose the more confident (and calibrated) agent, $prop = .77, z = 3.45, p < .001, BF_{alt} = 16.56$. In condition B, participants had a very slight nonsignificant preference for the overconfident (miscalibrated) agent over the calibrated one, $prop = .60, z = 1.12, p = .26, BF_{null} = 1.48$.

Participants' estimates for the accuracy and average confidence of the two agents are presented in Figure 4.27. Estimates were very similar for both advisers across all conditions, suggesting that they were strongly influenced by the betting phase statistics.

4.4.4 Discussion

Contrary to our predictions, participants did not make fine-grained optimal decisions about whose advice to take based on differences in the advisers' true probability of being correct on a given trial. This was not due to a failure to understand the task since performance in the control blocks conformed to predictions.

The failure to adapt decisions to the adviser profile is notable given that recalibration in this case was particularly easy (only a constant bias adjustment). Based on this, it is unlikely that more sophisticated recalibration could succeed, such as a difference in sensitivity. This is in

line with the findings of Pescetelli et al. (2016).

When comparing advisers who only differed in their confidence, participants were significantly more heavily influenced by the more confident adviser in condition A. This happened even though participants had observed the potential advisers over an extended number of trials and were presented with summary statistics to ensure that differences would be salient.

Results from condition A alone would lend support to the confidence heuristic. However, based on the lack of a similar pattern in condition B, we suggest that calibration was a mediating factor. Specifically, a more confident agent held more sway on participants when it was calibrated, but not when it was overconfident. It should be noted that this modulation of calibration was not actually beneficial in our task. The differential outcomes in the two conditions also suggest that the magnitude of the confidence manipulation was sufficient for participants to pick up on calibration differences, although, further confirmation with larger differences is needed.

However, we need to exert caution in the interpretation of the condition A and B differences given the null pattern in condition B was generated by very large inter-individual variability (that we could not explain based on the measure of task attentiveness). The source of the inter-individual variability is an interesting further direction. Here, differences were not linked to performance on the attention check task or the estimation task assessing learning.

Importantly, there was no indication in our experiment that either of the advisers had a motive to (or would incur any benefit from) strategic manipulations of their confidence. It is possible that in situations where the two advisers are competing (with each other or for the influence of the participant), people would exert more vigilance and results would more closely match our predictions.

The dissociation between the way in which participants made decisions about whose advice to take and how they computed their confidence in their judgments merits further attention. In the current design, there is some ambiguity between participants truly reporting confidence in their decision or confidence in the adviser. If it is indeed the case that confidence of advice takers is entirely determined by the confidence of the person whose advice was selected, and more

confident advisers are generally preferred (unless they are blatantly wrong), this can further amplify overconfidence as information is being circulated in social networks.

4.5 General Discussion

The first experiment showed that people have a preference for collaborating with others whose confidence is informative and commensurate with their accuracy. Findings disagreed with the confidence heuristic (Price & Stone, 2004) and support the calibration hypothesis (Pulford et al., 2018; Tenney et al., 2007).

In the second experiment, we asked whether participants act on the inferences about the metacognitive skills of advisers in an optimal way. We did not find evidence of recalibration, to the contrary, participants were more likely to choose a confident (and overall calibrated) adviser on trials where they were less likely to be correct than an overall underconfident adviser. Consistent with the calibration hypothesis, participants did not extend this undue influence to an overconfident adviser.

General preferences for collaborators in Experiment 2 matched the implicit decisions. Results converged with results of Experiment 1 in terms of a preference for a calibrated adviser over an underconfident one, but, in disagreement, there was no preference was found for the calibrated adviser when compared to an overconfident adviser. It seems likely that, like in Sah et al. (2013), there is a difference between explicitly verbalized preferences and actual implicit behavior.

A clear direction forward is to replicate the experiment with larger differences in confidence, to test the extent to which people will continue to take confidence statements at face value. Further, another replication can test if a cover story suggesting advisers may be attempting to exert influence over their audience may make participants more vigilant. Recent work casts down on the ability of humans to resist the strategic manipulations of advisers (Kurvers et al., 2021). Another factor that intuitively could be thought to have affected results is that it was conducted online and not in person, in a social context. However, Bower and Pulford

(2013) have found no differences between face-to-face and online advice utilization. Further, even the way in which confidence is presented, verbally or nonverbally, may impact results, for instance by varying the degree of plausible deniability associated with the communication channel (Tenney et al., 2019).

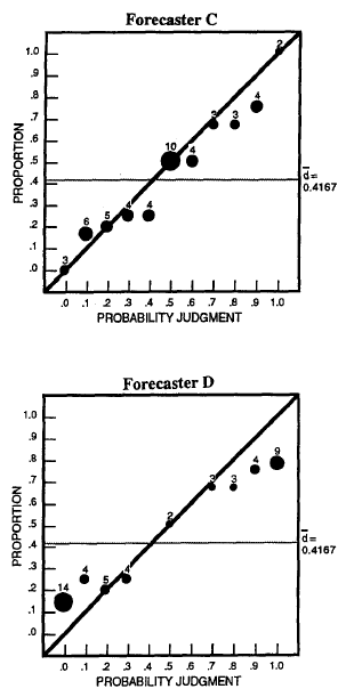
A further interesting direction is studying the inter-individual differences observed on this task. Previous work has already found that gender (Lee, 2005) is linked to adviser preferences, but there was little effect of personality differences (Pulford et al., 2018). A potential correlate for recalibration in the current design could be statistical learning.

More pertinently, we can ask what the present results entail for the original question - the ability of learners to choose reliable teachers and the extent to which teachers may use confidence to assess the progress of their learners. Results can be interpreted to mean that there is initial vigilance about who to learn from, but once an informant is deemed reliable, no further scrutiny is applied.

4.6 Supplementary Information

4.6.1 Experiment 1: Supplementary figures

Figure 4.17: Figure extracted from Yates et al. (1996): Calibration graphs of the forecast-outcome data used in the experiments

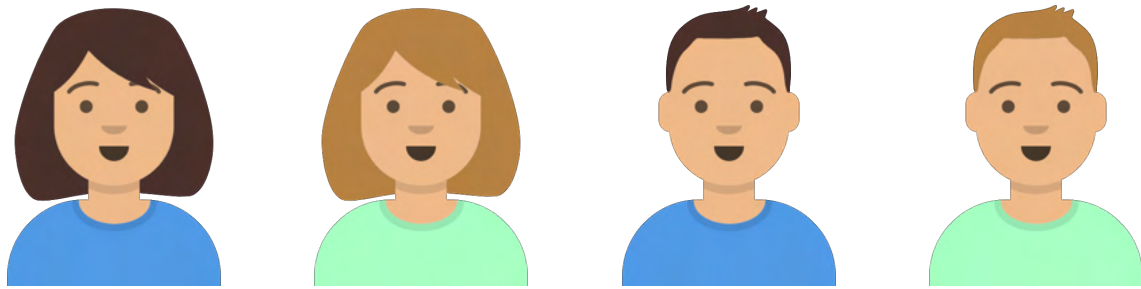


Stimuli

Figure 4.18: Example of a stimulus set used in Experiment 1 generated using code from Op de Beeck et al. (2001).

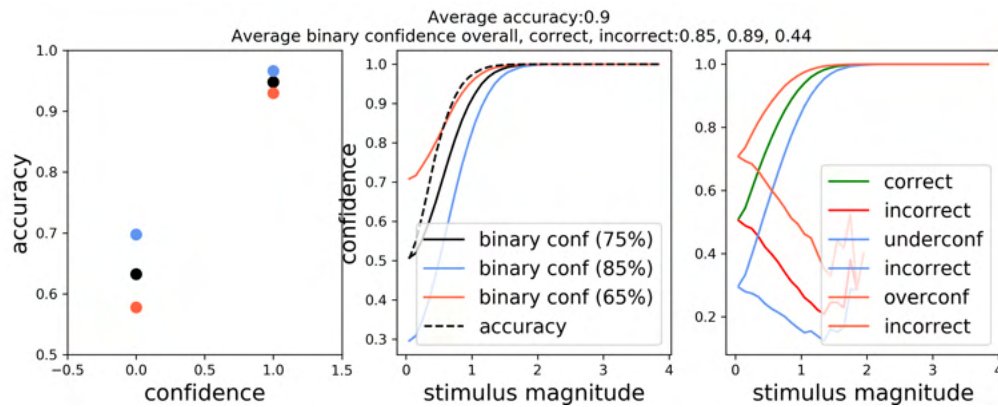


Figure 4.19: Avatars used in Experiment 1



Generation of confidence profiles

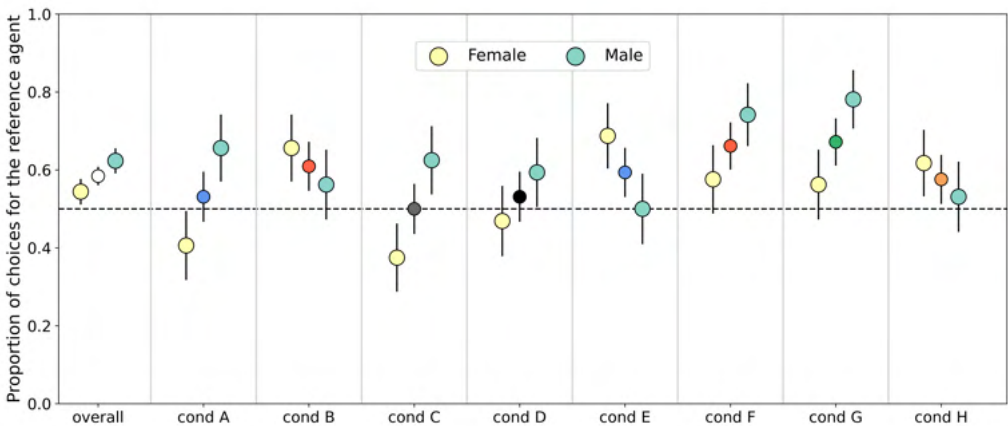
Figure 4.20: Origins of different confidence profiles



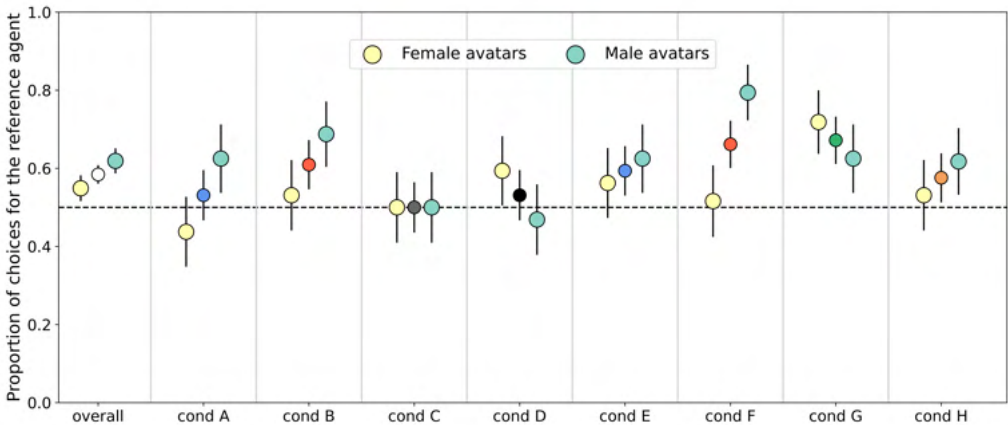
Experiment 1: Gender effects

Figure 4.21: The effect of gender on collaborator preferences.

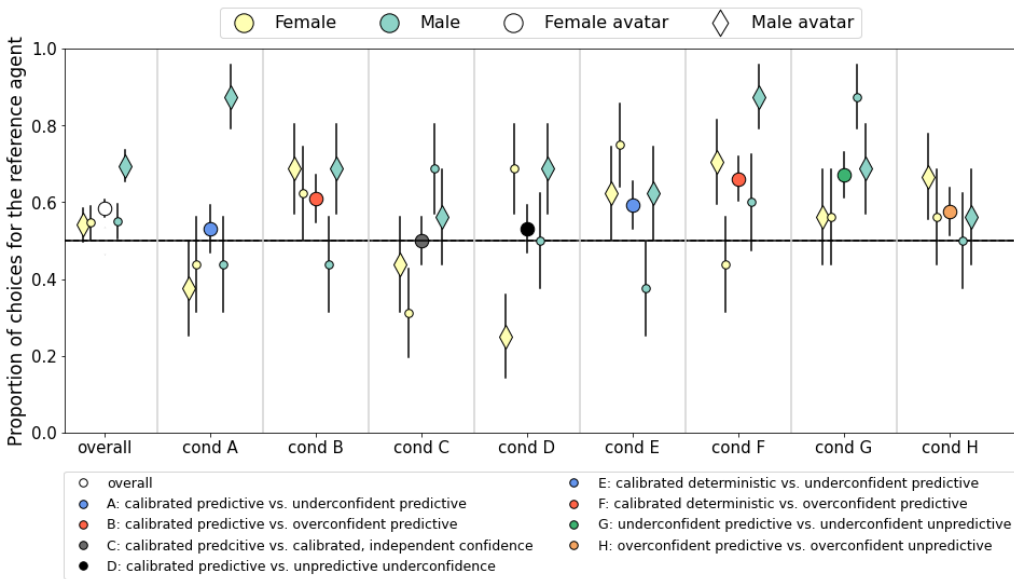
(a) Preferences for the calibrated agent as a function of the gender of the participant.



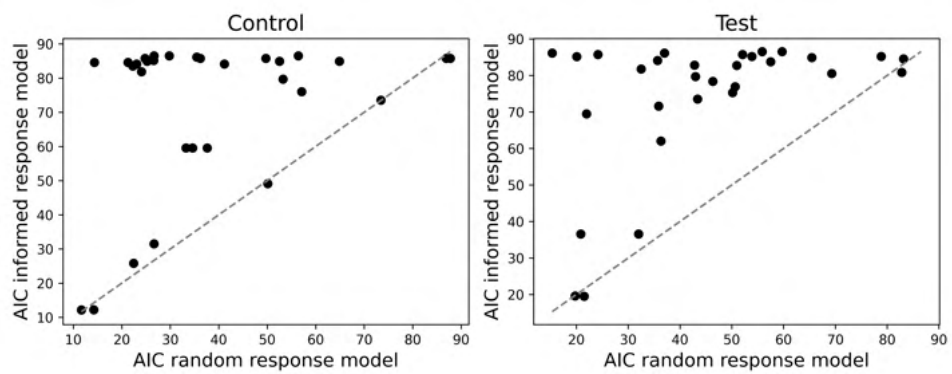
(b) Preferences for the calibrated agent as a function of the gender of the avatars.



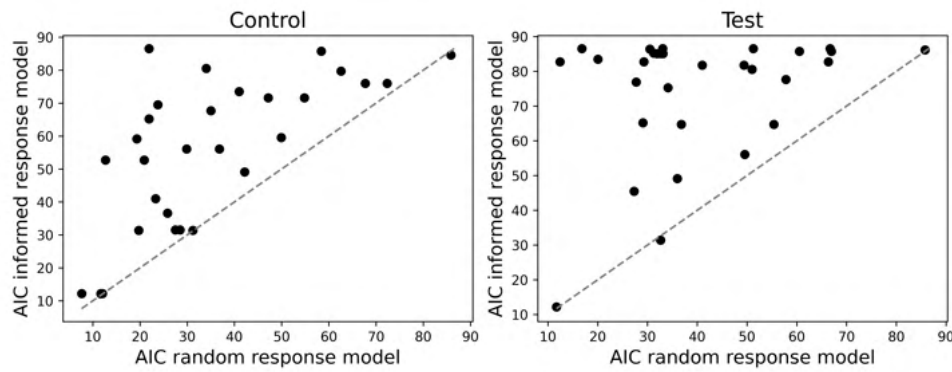
(c) Preferences for the calibrated agent as a function of the gender of the avatar pair and the gender of the participant.



4.6.2 Experiment 2: Supplementary figures



(a) Condition A



(b) Condition B

Figure 4.22: AIC comparison used for participant exclusions.

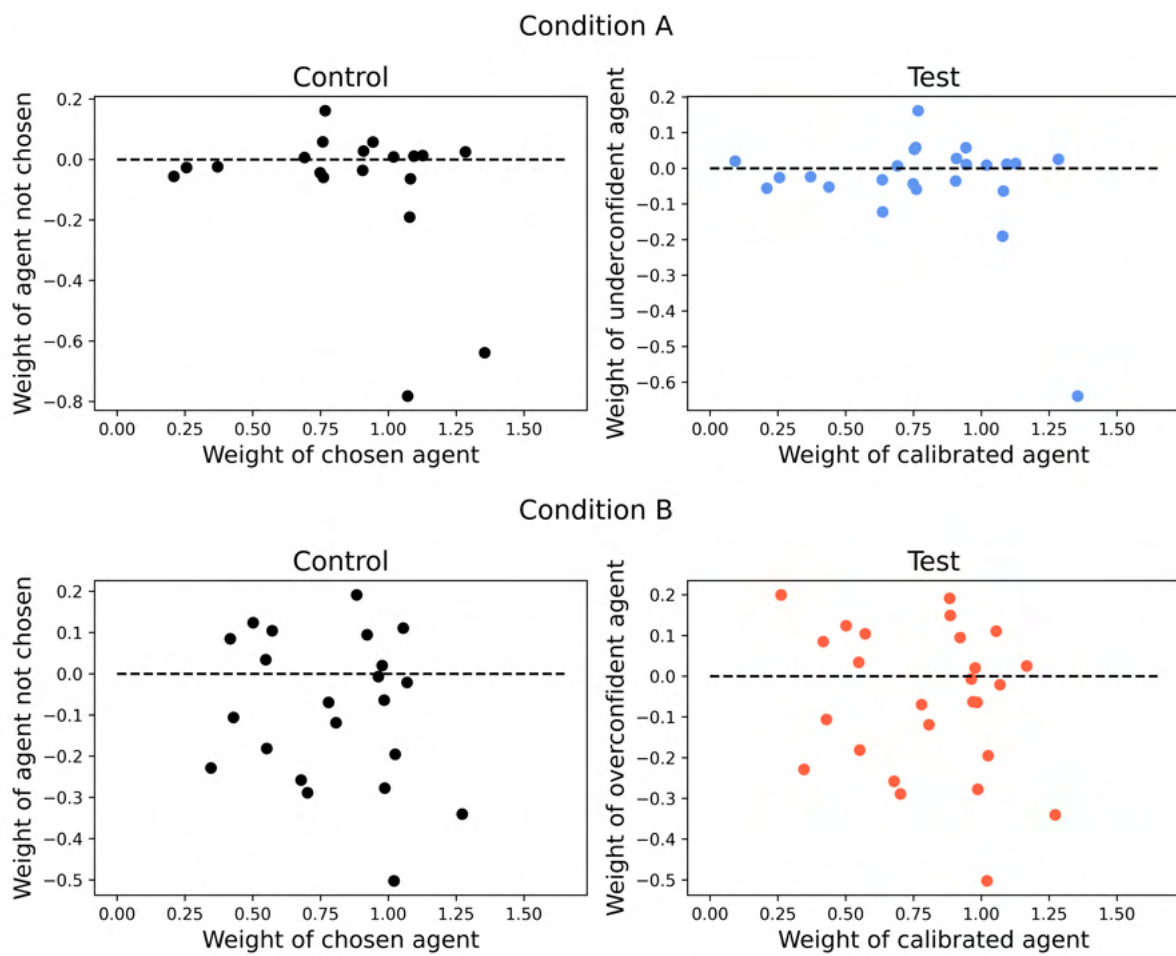


Figure 4.23: Regression weights for predicting participants' confidence based on the confidence of advisers. Each dot is a participant.

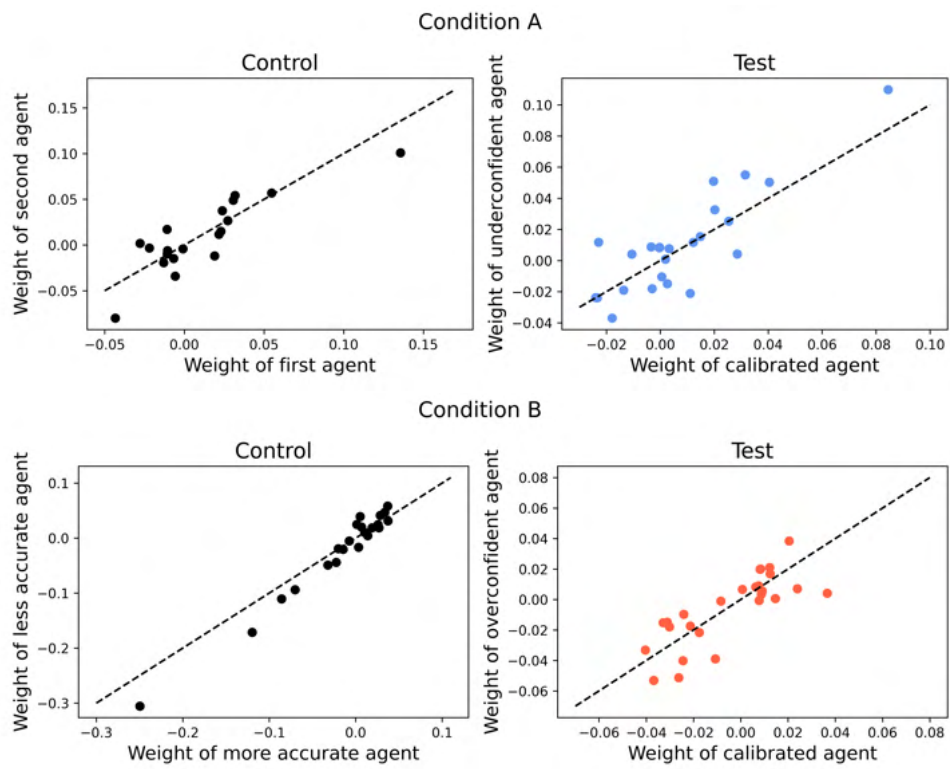


Figure 4.24: Fitted weights from logistic regression of participant decisions on the two advisers' continuous judgments (signed confidence). Each dot is a participant.

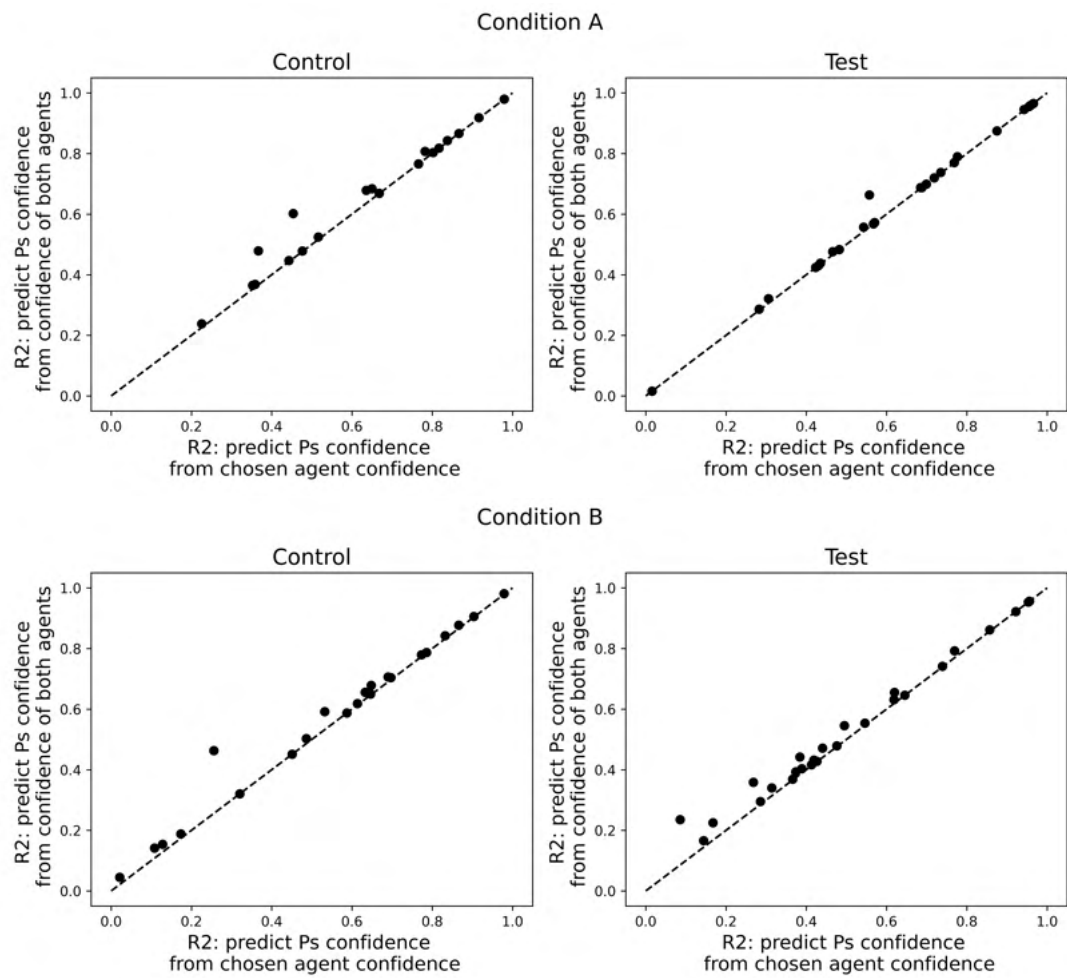


Figure 4.25: Variance explained(R^2) for models of participants' confidence. Each dot represents a participant.

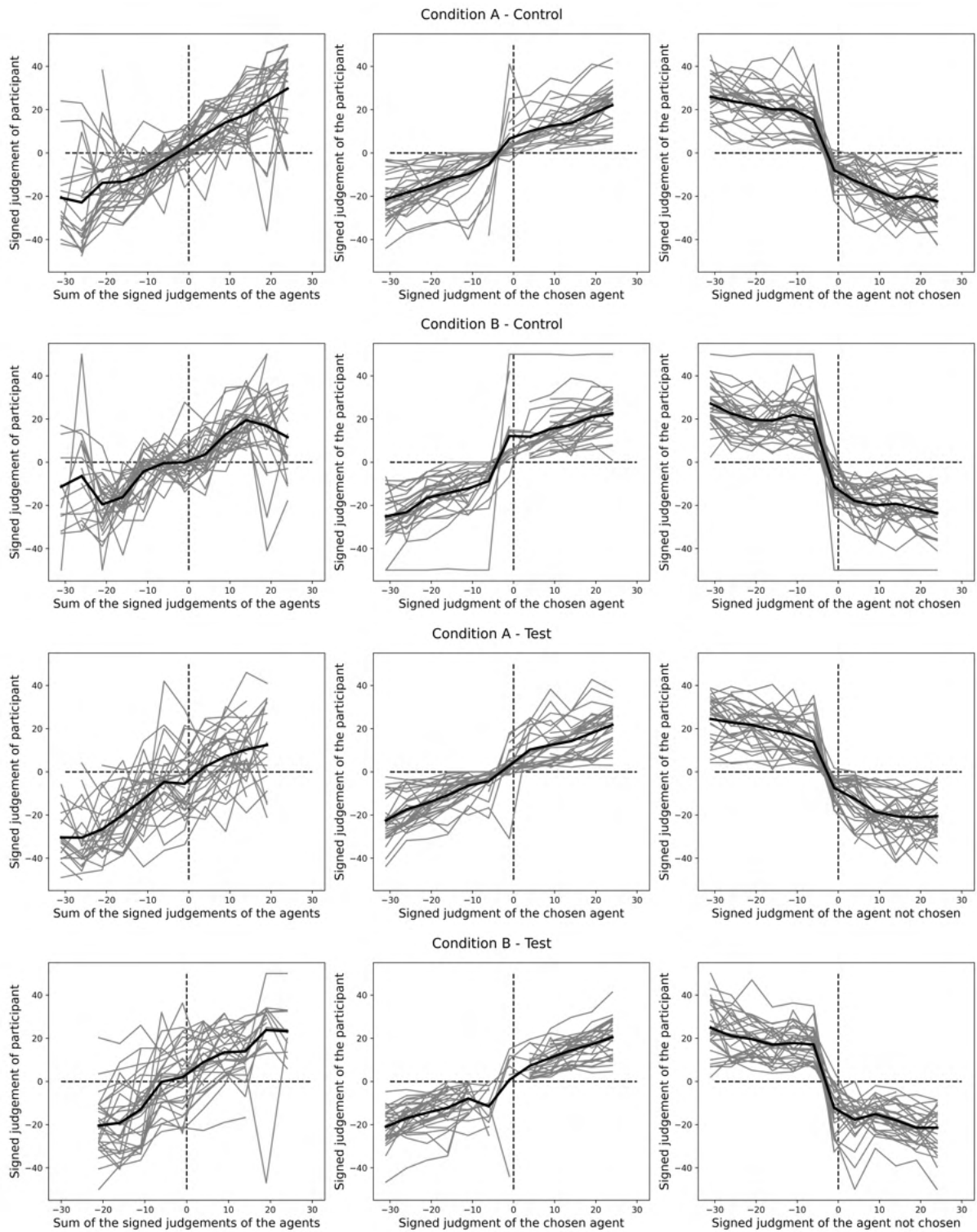


Figure 4.26: Participants' judgments as a function of adviser's judgments. Each line corresponds to a participant.

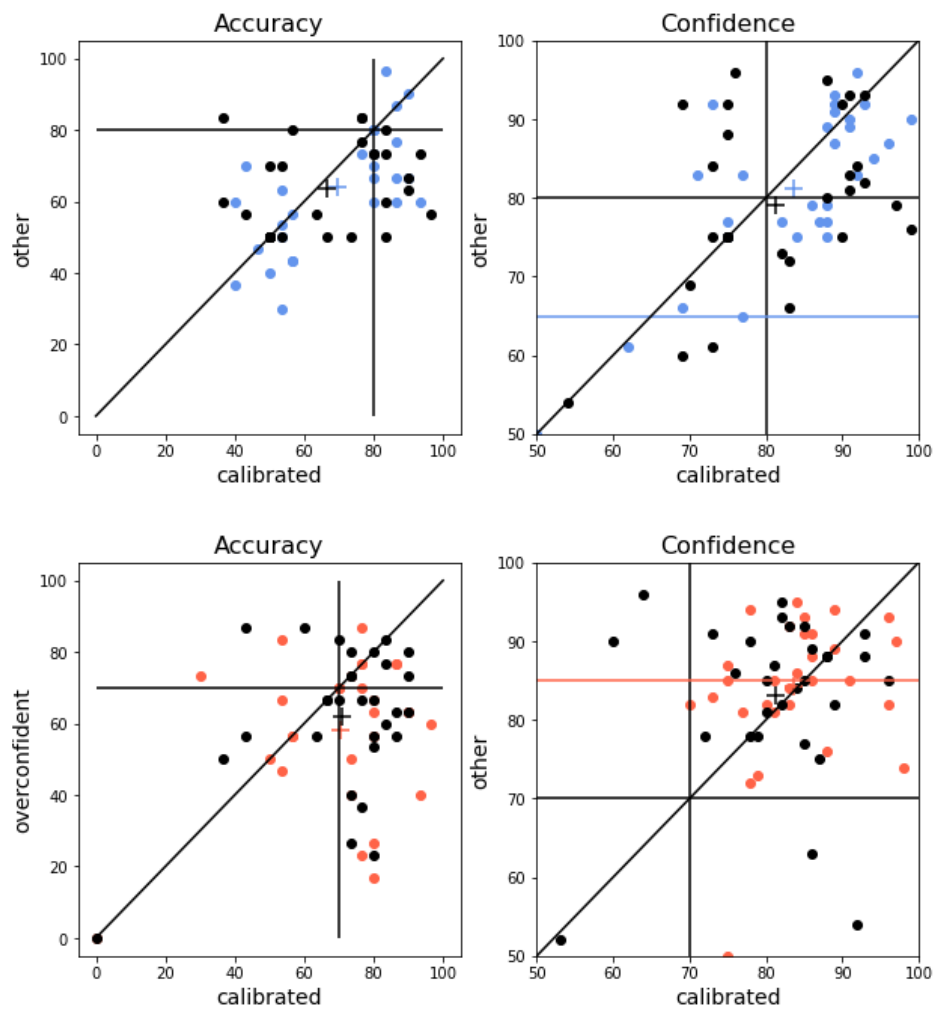


Figure 4.27: Participant estimates for the two advisers' accuracy and confidence. Participants did not see numerical values on the scales, the half slider length was rescaled for plotting to the 50-100 range. Each dot is a participant. Crosses represent sample averages. Vertical and horizontal lines are the ground truth values.

4.6.3 Experiment 2: Pilot

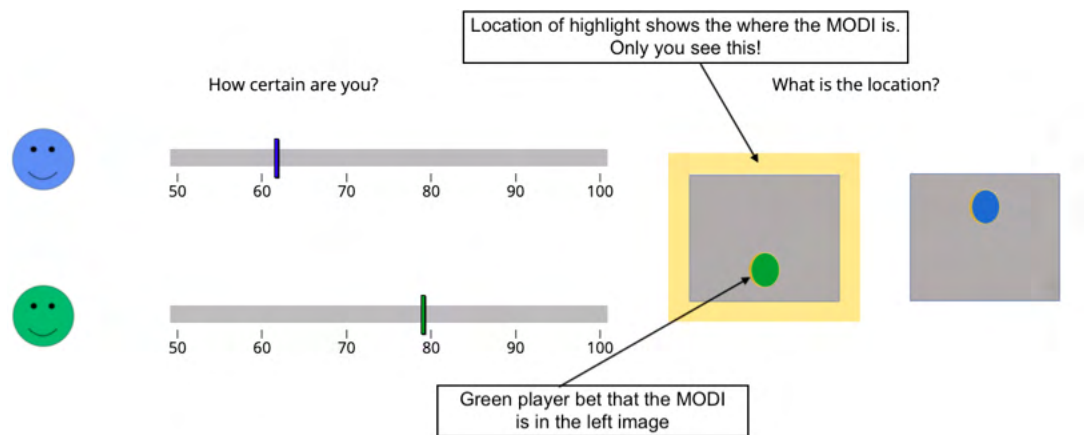


Figure 4.28: Observation phase of pilot.

The pilot experiment had the same structure as Experiment 2, with some differences in the visual presentation (see [Figure 4.28](#)) of the observation trials and the way in which participant responses were made. During observation, the choices of the advisers and their confidence were presented separately. Scales showed confidence ratings from 50% to 100%, marking every 10 point increment and had no verbal labels (although the scale was extensively explained at the beginning of the experiment). Second, no additional information was provided to reinforce the accuracy and confidence levels of the advisers. During the betting phase, participants first made a 2AFC about which adviser's suggestion to take, and subsequently stated their confidence in their decision. Both test conditions were piloted, but only one of the control conditions, specifically the one comparing two identical advisers.

32 participants (16 per condition) took part in the pilot. Main results are presented in [Figure 4.29](#). There was considerable inter-individual variability in boundaries, including in the control condition. This is indicative of the fact that perception of the differences between agents was very noisy and lead to the simplification of the visual presentation of the advisers during observation and the addition of the summary presentation. While there is a high variability, qualitatively results were taken to support the recalibration prediction as boundaries were biased in the predicted direction.

In light of the results of Experiment 2, it is worth considering what is responsible for the apparent differences in outcomes. The changes made were intended to make the task easier to follow and more intuitive for participants. The changes had the intended outcome as they reduced the noise in the control conditions. At the same time, they also led to a change in direction in condition A. We speculate that the use of verbal labels instead of a numbered scale was responsible for this difference, by driving home the meaning of the values presented on the scale, as opposed to treating accuracy and confidence like any other variables that may be statistically associated.

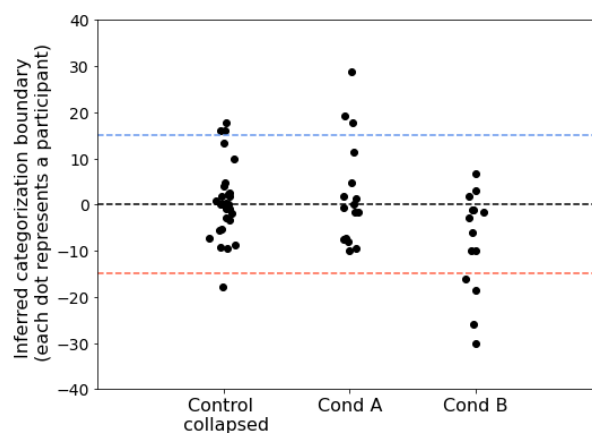


Figure 4.29: Participants' individual boundaries, inferred by regressing binary choices on the difference in adviser confidence. Dashed horizontal lines represent the predictions based on recalibration of confidence.

Chapter 5: General Discussion

5.1 Conclusions

A large part of human learning occurs in explicitly pedagogical settings or in what can be broadly construed as communicative events which entail purposeful information transmission. The prevalence of such settings can be explained by the fact that they tend to be conducive to more efficient learning. In fact, it has been proposed that part of the puzzle of how young children learn so much so quickly lies in their ability to capitalize on pedagogical situations (Gergely et al., [2002](#)).

The behavioral literature to date has been heavily unbalanced, focusing more on learning compared to teaching. Investigating teaching behaviors is crucial given both the practical consequences for parents and educators, and the theoretical implications regarding the explanatory power of teaching for the incredible induction abilities of humans.

An important way in which people teach is by choosing good examples that would enable learners to make more robust inferences faster. This is also an interesting task to study because of its inherent complexity. To be (near-)optimal, teachers need to have a good approximation of the generative model of the task and, additionally, an accurate model of the learner (i.e., what the learners know and how new information changes their beliefs). Further, teachers must implement a data sampling algorithm (e.g, rational pedagogical sampling) that produces the example set leading to the best learning outcome in light of this knowledge, and, in doing so, must overcome any implicit sampling biases along the way.

In Chapter 2 we found that humans are able to perform contextually dependent purpose-

ful sampling, which could be distinguished from generating data through an uninformative (random) process. The results followed the predictions of a normative model Shafto et al. (2014), which can be used to describe the optimal outcome of collaboration in teacher-learner dyads. Results were also consistent with reliance on simpler strategies such as summary statistics matching or conveying exemplars as varied as possible. These strategies could be useful approximations to the optimal solutions. We argue that the main benefit of the Shafto et al. (2014) pedagogical sampling model is that it offers a way to derive predictions about outcome optimality; instead of providing a psychologically realistic account of human teaching.

However, we also observed suboptimal performance on a conceptually simple task (one-dimensional boundary teaching), which highlights the fact that the assumption of rationality does not do enough work in practical pedagogy. This result is meaningful especially since on this task it was sufficient to teach a ‘literal learner’. That is, no additional assumptions were required or would aid the learner to identify the hypothesis to be taught based on the examples provided.

Therefore, we suggest that teaching-by-examples breaks down in various circumstances: when the level of abstraction of the task is low, when more complex representations of the task are built (even if they are extraneous to the teaching problem), and when there is uncertainty about how a potential learner would reason about the evidence.

Focusing on the matching learning process, we found that learners were sensitive to the generative process of sequentially presented data, a finding that resonates with the rich literature on perceptual learning. Our specific goal was to contrast how humans produce estimates based on i.i.d. sequentially presented series with strongly autocorrelated series. We were interested in autocorrelation since it is ubiquitous in serial inputs that originate both from the natural environment and from other humans. Given the prevalence of correlations, we had a strong expectation that participants would be able to adapt to the statistical structure to improve their performance in an estimation task. In fact, it has been shown that even in i.i.d. series, illusory positive serial correlations are detected. This leads to a hot-hand bias (Blanchard et al., 2014), which is potentially the result of an ecologically rational adaptation.

In the case of autocorrelated inputs, the optimal pattern for computing the mean of the series entails symmetric overweighting of the first and last samples in the series, a ‘primacy and recency’ effect. This would have provided a novel normative explanation for serial effects (especially for early sample overweighting). However, we did not find convincing evidence in favor of this pattern at the individual level. It is possible that the current task did not sufficiently incentivize the use of the optimal strategy (e.g., by imposing a cost function), and further experiments may provide evidence to support our prediction.

At a glance, the results presented in this chapter paint a pessimistic picture about the feasibility of optimal teaching and learning from teaching. However, we propose that in real world scenarios teachers can rely on varied sources of information to boost their performance.

Therefore, in Chapter 3 we asked how teachers may bypass one of the most difficult challenges of teaching, inferring a model of their learner, by building such a model based on their own learning experience. We found support for the fact that learning experience improves subsequent teaching-by-examples performance in two different experiments. Moreover, we showed that active learning is more useful than passive learning by using yoked designs in which active learners were paired with passive learners. Why does active learning help? The teaching task shares considerable similarity with active learning - a sample needs to be produced and evaluated by some criterion: expected utility for learning (e.g. uncertainty reduction) in the case of active learning and in the case of teaching, magnitude of learning achieved towards ground truth.

The first experiment used a simple one dimensional boundary teaching task, with few stimuli, where the boundary was explicitly known and visually present throughout the entire experiment. Arguably, this was an insight task, as the optimal strategy, once discovered, was easy to describe verbally (‘choose stimuli directly on either side of the boundary’) and transfer explicitly. However, in the second experiment, a controlled setup was used in which learning was implicit and the change in representation could be quantified. The category learning task, adapted from Markant and Gureckis (2014), was extended in time and implicit, as participants generally could not report back the structure of the categories they had learned. This is the first

teaching experiment, to our knowledge, where the teacher has to first acquire a novel concept by learning that is extended in time, and allows for studying the relationship between latents describing the representation extracted and subsequent teaching.

The improvements from prior learning experience occurred when the categorization task was more difficult, for information-integration categories as opposed to rule-based ones. This is in line with predictions, as the hypothesis space for the two-dimensional classification task does not match participants' expectations. Overall, we found that participants were overly cautious in their example generation - even participants who had acquired precise knowledge of the location of the boundary were not offering examples close to the boundary that would reduce most uncertainty for the learner. However, this is in line with aiming to reduce mistakes for highly noisy learners. Moreover, while active learning did improve teacher performance according to objective learning metrics, we found that the examples provided by teachers were consistent with strong sampling. We would like to confirm in the future whether the examples chosen by teachers indeed lead to improved performance for naïve learners on this task compared to strong sampling.

The results of Chapter 3 were robust, and replicated, but the magnitude of the improvement in teaching from active learning experience was small. In other words, there is scope for training teachers, but it is possible that teaching the teacher will not yield the sizable effects that are needed to make a real difference for learners.

In Chapter 4, we focused on one way in which learners may cope with suboptimal teachers. Specifically, we proposed that learners may use confidence to distinguish reliable teachers from teachers who are manipulative or have no self-knowledge. Further, teachers may use confidence to assess the progress of their learners.

The first experiment revealed that participants were able to spot informants who were under or over-estimating their chances of success and preferred calibrated and informative advisers as future collaborators. However, it should be noted that effect sizes were modest, presumably due to the general difficulty level of the task which required extended passive observation to determine quantitative differences in the calibration profiles based on conditional probabilities

of being correct given being confident or not. It would be particularly interesting to test whether if a competitive scenario between the two advisers is introduced to participants, they would increase their vigilance and show more extreme preferences. Lastly, we (unexpectedly) found a host of gender effects that warrant further exploration with a targeted design in light of their magnitude relative to the magnitude of the predicted effects.

Despite the pattern observed above for collaborator preferences, in a subsequent experiment, participants could not leverage that information in order to make optimal choices about who to learn from on a trial-by-trial basis. The prediction of the experiment was that participants who learned the functional mapping between confidence and accuracy would make predictions for the expected accuracy of an adviser when only their confidence judgements are shown. Thus, we predicted that a recalibrated confidence would be used trial-by-trial when choosing between two known advisers, as opposed to explicit confidence judgements.

Reassuringly, selective learning decisions were optimal when differences in competence were introduced between advisers. A more confident agent, when calibrated, held undue influence on participant decisions. However, a more confident but overconfident agent did not. This is an interesting result, in line with previous work highlighting the use of calibration as a credibility signal (Tenney et al., 2011). We speculate that results are indicative of the fact that instead of using the relationship between the confidence and accuracy of an adviser to make detailed predictions about individual learning instances, which is cognitively very taxing, humans make use of calibration information to formulate priors about who is credible that are subsequently used to inform decision making.

What can we conclude based on the cumulative results presented here? As expected, teaching by giving examples was a hard problem even in well-defined and low-complexity tasks, and even after prior experience with it. Namely, in our experiments teachers struggled to create learning sets which eliminated sufficient amounts of uncertainty for the learner. However, we speculate that the strategies used by teachers were adapted to learners who are themselves limited. For instance, relying on teaching of extreme examples from two categories, that teachers were very confident about, reduces the risk that a particularly noisy learner would accidentally

mislabel samples in its vicinity. On flip side, avoiding misleading learners is also critical if teachers want to assure that learners continue to follow their advice. This has been shown in Chapter 4 where correct and confident informants were preferred as collaborators.

In general, it appears that the future steps for understanding learning should include building theoretical models and experimental paradigms which have an increased degree of psychological realism and address how plausibly constrained learners and teachers can interact.

5.2 General limitations and further directions

The main limitation of the work presented and that we intend to pursue in the future is that studying if and how teachers adapt on the fly to the learner in closed-loop experiments. There are several ways in which current experiments can be modified for **interactive settings**. The second experiment of Chapter 3 was designed specifically as a precursor of a closed-loop experiment in which the teacher observes a learner in real time and is sporadically allowed to offer examples to the learner. Teaching sets in this case are not expected to converge across the entire sample, as teachers ought to flexibly adapt to the idiosyncrasies of their learners. An interesting question to explore in this setup is whether teachers are adapting the number of interventions and distance to the boundaries of the examples to the (inferred) uncertainty of the learner around the boundary.

Further, we plan to conduct a manipulation of the mode of learning (active or passive) to test whether teachers who observe active learners are able to make more precise inferences about the learner's current state of knowledge and are, therefore, better able to target their teaching examples. In other words, do learners indirectly benefit from active exploration by making their mental states easier to read? This rests on an implicit assumption on the part of the teachers that active learners are trying to minimize their uncertainty with respect to the boundary location. This assumption seems plausible as form of informational naïve utility calculus (Jara-Ettinger et al., 2016).

It is also interesting to explore whether closed-loop dynamics would improve teacher per-

formance. On the one hand, access to immediate feedback from the learner in the form of questions or checks of their comprehension, would facilitate teaching by making salient which aspects of the teaching were successful or not. One direction that we believe to be particularly promising for both theoretical and empirical exploration is the use of learners' confidence judgments to fine tune teaching. Work in education has shown that often students can answer difficult questions correctly without having a deep understanding of the justifications behind it. Confidence judgments have the potential to distinguish lucky correct answers from justified correct answers. Further, confident wrong answers might suggest an improper understanding of the answer and the need for extra clarification.

Second, human **teaching takes multiple forms**, from physical demonstrations to aid motor skill acquisition or tool use, to complex verbal instruction such as seen in formal education. What is the place of teaching-by-example in this ecology?

While in the current studies we looked only at example generation, the computational framework is general and can apply to any task for which a hypothesis space and a set of actions can be defined. For instance, the same principles have been used to show how teachers should choose an intervention to identify a causal net or to which verbal labels to use to identify a referent. It remains an open question whether there are meaningful differences in the ability to apply the iterated reasoning of pedagogical sampling to these different domains. Alternative computational frameworks can potentially be used to explain different aspects of teaching. For instance, partially observable Markov decision processes are more suited to model teaching as a planning problem Rafferty et al. (2011) and F. Wang (2014).

Related this point, there is a strong reliance on **categorizations tasks** throughout the chapters of the thesis. These were preferred as they are a natural match for studying teaching through example giving and make it easier to quantify behavior. Categorization also has a long tradition in the cognitive science literature and is, therefore, fairly well understood both behaviorally and computationally. However, a valid critique can be brought about the ecological validity of such tasks to assess teaching. Categorization, understood as a means of structuring the world in meaningful ways and solidifying representations, is a core ability and forms the

basis for generalization.

However, it is likely the case that a large amount of human teaching of categorization is not done by example giving (while this undoubtedly is part and parcel of it), but by verbally providing a categorization rule or describing relevant features. On the other hand, there is certainly a very large proportion of categorization structures that cannot be described effectively through comprehensible rules (as opposed to mathematical descriptions) that would facilitate learning. For example, a recent study by Rosedahl et al. (2021) has shown that for information-integration category structures, explicit verbal instruction does not aid learning.

Third, while teaching is a continuous part of the human experience and adult-to-adult teaching is common, the paradigmatic setting to study teaching is set early in development, where the **knowledge asymmetry** between the teacher and the learner is the largest. Unfortunately, this is also the situation in which the rational pedagogy model fairs the worst as a consequence of the common ground violation (Yang & Shafto, 2017a).

The **teaching goals** in all the tasks examined were well defined, but this may not be the case in real-life situations. We need to start asking questions about how effective teaching is beyond generalization in classifications tasks, moving into more ecological transfer situations. For example, is it possible to teach students how to generate ‘better’ hypotheses on their own or a sense for what makes a new hypothesis ‘good’? Another important skill to be transmitted is how to structure the learning problems into (ordered) subtasks in a way that is conducive to finding a solution. There are countless further goals that teachers can have in mind for their learners: how to stay motivated, estimate which learning challenges are too easy or too difficult to attempt.

Moreover, an important part of learning is to gain an understanding of what strategies are most successful in solving particular types of problems. This can be thought of as a metareasoning problem where a learner needs to choose the best algorithm in terms of effectiveness but also efficiency given the data it needs to be applied to, time and computational costs (Lieder et al., 2014). Certainly this is what most teachers aim to achieve, namely transmitting a toolbox of cognitive strategies that can be deployed adaptively.

References

- Adler, W. T., & Ma, W. J. (2017, November 13). *Limitations of proposed signatures of bayesian confidence* (preprint). Neuroscience. <https://doi.org/10.1101/218222>
- Aitchison, L., Bang, D., Bahrami, B., & Latham, P. E. (2015). Doubly bayesian analysis of confidence in perceptual decision-making (A. A. Faisal, Ed.). *PLOS Computational Biology*, 11(10), e1004519. <https://doi.org/10.1371/journal.pcbi.1004519>
- Anderson, J. (1990). *The adaptive character of thought* (1st ed.) <https://doi.org/https://doi.org/10.4324/9780203771730>
- Ashby, F. G., & Maddox, W. (2005). Human category learning. *Annual review of psychology*, 56, 149–178.
- Ashby, F. G., Queller, S., & Berretty, P. M. (1999). On the dominance of unidimensional rules in unsupervised categorization. *Perception & Psychophysics*, 61, 1178–1199.
- Ashby, F. G., Maddox, W. T., & Bohil, C. J. (2002). Observational versus feedback training in rule-based and information-integration category learning. [Place: US Publisher: Psychonomic Society]. *Memory & Cognition*, 30(5), 666–677. <https://doi.org/10.3758/BF03196423>
- 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. <https://doi.org/10.1080/713755719>

- Bahrami, B., Olsen, K., Latham, P. E., Roepstorff, A., Rees, G., & Frith, C. D. (2010a). Optimally interacting minds. *Science*, 329(5995), 1081–1085. <https://doi.org/10.1126/science.1185718>
- Bahrami, B., Olsen, K., Latham, P. E., Roepstorff, A., Rees, G., & Frith, C. D. (2010b). Optimally interacting minds. *Science*, 329(5995), 1081–1085. <https://doi.org/10.1126/science.1185718>
- Bahrami, B., Olsen, K., Bang, D., Roepstorff, A., Rees, G., & Frith, C. (2012). What failure in collective decision-making tells us about metacognition. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 367(1594), 1350–1365. <https://doi.org/10.1098/rstb.2011.0420>
- Baker, C. L., Jara-Ettinger, J., Saxe, R., & Tenenbaum, J. B. (2017). Rational quantitative attribution of beliefs, desires and percepts in human mentalizing. *Nature Human Behaviour*, 1(4).
- Bang, D., Aitchison, L., Moran, R., Herce Castanon, S., Rafiee, B., Mahmoodi, A., Lau, J. Y. F., Latham, P. E., Bahrami, B., & Summerfield, C. (2017). Confidence matching in group decision-making. *Nature Human Behaviour*, 1(6), 0117. <https://doi.org/10.1038/s41562-017-0117>
- Bargh, J. A., & Schul, Y. (1980). On the cognitive benefits of teaching. [Place: US Publisher: American Psychological Association]. *Journal of Educational Psychology*, 72(5), 593–604. <https://doi.org/10.1037/0022-0663.72.5.593>
- Barthelmé, S., & Mamassian, P. (2009). Evaluation of objective uncertainty in the visual system (K. Kording, Ed.). *PLoS Computational Biology*, 5(9), e1000504. <https://doi.org/10.1371/journal.pcbi.1000504>
- Bass, I., Bonawitz, E. B., Shafto, P., Ramanajan, D., Gopnik, A., & Wellan, H. (2017). I know what you need to know: Children’s developing theory of mind and pedagogical evidence selection.

- Bass, I., Shafto, P., & Gopnik, A. (2017). I know what you need to know: Children's developing theory of mind and pedagogical evidence selection. *Proceedings of the 39th Annual Conference of the Cognitive Science Society*, 6.
- Behrens, T. E. J., Hunt, L. T., Woolrich, M. W., & Rushworth, M. F. S. (2008). Associative learning of social value. *Nature*, 456(7219), 245–249. <https://doi.org/10.1038/nature07538>
- Bengio, Y., Louradour, J., Collobert, R., & Weston, J. (2009). Curriculum learning, 41–48.
- Birch, S. A. J., Akmal, N., & Frampton, K. L. (2010). Two-year-olds are vigilant of others non-verbal cues to credibility. [Place: United Kingdom Publisher: Wiley-Blackwell Publishing Ltd.]. *Developmental Science*, 13(2), 363–369. <https://doi.org/10.1111/j.1467-7687.2009.00906.x>
- Blanchard, T. C., Wilke, A., & Hayden, B. Y. (2014). Hot-hand bias in rhesus monkeys. [Place: US Publisher: American Psychological Association]. *Journal of Experimental Psychology: Animal Learning and Cognition*, 40(3), 280–286. <https://doi.org/10.1037/xan0000033>
- Bonawitz, E., Shafto, P., Gweon, H., Goodman, N. D., Spelke, E., & Schulz, L. (2011a). The double-edged sword of pedagogy: Instruction limits spontaneous exploration and discovery. *Cognition*, 120(3).
- Bonawitz, E., Shafto, P., Gweon, H., Goodman, N. D., Spelke, E., & Schulz, L. (2011b). The double-edged sword of pedagogy: Instruction limits spontaneous exploration and discovery. *Cognition*, 120(3), 322–330.
- Bower, S. D., & Pulford, B. D. (2013). Utilization of advice from face-to-face and internet-mediated advisors. *Journal of Technology in Human Services*, 31(4), 304–320. <https://doi.org/10.1080/15228835.2013.855697>
- Brainard, D. H. (1997). The psychophysics toolbox. *Spatial Vision*, 10, 433–436.
- Bruner, S., J. (1961). The act of discovery. *Harvard Educational Review*, 31, 21–32.
- Burson, K. A., Larrick, R. P., & Klayman, J. (2006). Skilled or unskilled, but still unaware of it: How perceptions of difficulty drive miscalibration in relative comparisons. *Journal*

- of Personality and Social Psychology*, 90(1), 60–77. <https://doi.org/10.1037/0022-3514.90.1.60>
- Cakmak, M., & Thomaz, A. L. (2010). Optimality of human teachers for robot learners. *2010 IEEE 9th International Conference on Development and Learning*, 64–69.
- Chi, M. T. H., Siler, S. A., & Jeong, H. (2004). Can tutors monitor students' understanding accurately? cognition and instruction, 22(3).
- Cohen, J. D., McClure, S. M., & Yu, A. J. (2007). Should I stay or should I go? How the human brain manages the trade-off between exploitation and exploration. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, 362(1481).
- Corriveau, K., & Harris, P. L. (2009). Choosing your informant: Weighing familiarity and recent accuracy. *Developmental Science*, 12(3), 426–437.
- Csibra, G., & Gergely, G. (2011). Natural pedagogy as evolutionary adaptation. *In Culture Evolves*, 377–392.
- Csibra, G. (2010). Recognizing communicative intentions in infancy. *Mind & Language*, 25(2), 141–168. <https://doi.org/https://doi.org/10.1111/j.1468-0017.2009.01384.x>
- Csibra, G., & Gergely, G. (2006). Social learning and social cognition: The case for pedagogy.
- Csibra, G., & Gergely, G. (2009). Natural pedagogy. *Trends in Cognitive Sciences*, 13(4), 148–153.
- Cutler, B. L., Penrod, S. D., & Stuve, T. E. (1988). Juror decision making in eyewitness identification cases. *Law and Human Behavior*, 12(1), 41–55. <https://doi.org/10.1007/BF01064273>
- Duñabeitia, J. A., Crepaldi, D., Meyer, A. S., New, B., Pliatsikas, C., Smolka, E., & Brysbaert, M. (2018). MultiPic: A standardized set of 750 drawings with norms for six European languages. *Quarterly Journal of Experimental Psychology*, 71(4), 808–816.
- Ebbinghaus, H. (1850-1909). Memory: A contribution to experimental psychology. *H. Ruger, & C. Bussenius, Trans., 1913, New York, NY: Teachers College.*
- EHEA, R. C. (2020). <https://rm.coe.int/rome-ministerial-communicue-19-11-20/1680a07857>

- Ernst, M. O., & Banks, M. S. (2002). Humans integrate visual and haptic information in a statistically optimal fashion. *Nature*, 415(6870), 429–433. <https://doi.org/10.1038/415429a>
- Fiser, J., & Aslin, R. N. (2001). Unsupervised statistical learning of higher-order spatial structures from visual scenes [PMID: 11760138]. *Psychological Science*, 12(6), 499–504. <https://doi.org/10.1111/1467-9280.00392>
- Fiser, J., Berkes, P., Orbán, G., & Lengyel, M. (2010). Statistically optimal perception and learning: From behavior to neural representations. *Trends in Cognitive Sciences*, 14(3), 119–130. <https://doi.org/10.1016/j.tics.2010.01.003>
- Fleming, S. M., Weil, R. S., Nagy, Z., Dolan, R. J., & Rees, G. (2010). Relating introspective accuracy to individual differences in brain structure. *Science*, 329(5998), 1541–1543. <https://doi.org/10.1126/science.1191883>
- Frank, M. C., & Goodman, N. D. (2012). Predicting pragmatic reasoning in language games. *Science*, 336(6084), 998–998. <https://doi.org/10.1126/science.1218633>
- Futó, J., Téglás, E., Csibra, G., & Gergely, G. (2010). Communicative function demonstration induces kind-based artifact representation in preverbal infants. *Cognition*, 117(1), 1–8. <https://doi.org/https://doi.org/10.1016/j.cognition.2010.06.003>
- Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (2003). *Bayesian data analysis*. 2nd ed. CRC Press, London. MR2027492.
- Gergely, G., Bekkering, H., & Király, I. (2002). Rational imitation in preverbal infants. *Nature*, 415(6873), 755–755. <https://doi.org/10.1038/415755a>
- Gershman, S. J., Horvitz, E. J., & Tenenbaum, J. B. (2015). Computational rationality: A converging paradigm for intelligence in brains, minds, and machines. *Science*, 349(6245).
- Gilden, D. L., Thornton, T., & Mallon, M. W. (1995). 1/i_lf_i/i_l noise in human cognition. *Science*, 267(5205), 1837–1839. <https://doi.org/10.1126/science.7892611>
- Goldman, S. A., & Kearns, M. J. (1995). On the complexity of teaching. *Journal of Computer and System Sciences*, 50(1), 20–31.

- Goldman, S. A., & Mathias, H. D. (1996). Teaching a smarter learner. *Journal of Computer and System Sciences*, 52(2), 255–267.
- Goodman, N. D., Ullman, T. D., & Tenenbaum, J. B. (2011). Learning a theory of causality. [Place: US Publisher: American Psychological Association]. *Psychological Review*, 118(1), 110–119. <https://doi.org/10.1037/a0021336>
- Gopnik, A., & Bonawitz, E. (2015). Bayesian models of child development. *Wiley Interdisciplinary Reviews. Cognitive Science*, 6(2), 75–86.
- Gureckis, T. M., & Markant, D. B. (2012). Self-directed learning: A cognitive and computational perspective. *Perspectives on Psychological Science: A Journal of the Association for Psychological Science*, 7(5), 464–481.
- Gweon, H., Chu, V., & Schultz, L. E. (2014). To give a fish or to teach how to fish? Children weigh costs and benefits in considering what information to transmit.
- Gweon, H., Shafto, P., & Schulz, L. (2014). Considering prior knowledge and the cost of information both in learning from and teaching others. *Proceedings of the 36th Annual Conference of the Cognitive Science Society*.
- Gweon, H., Tenenbaum, J. B., & Schulz, L. E. (2010). Infants consider both the sample and the sampling process in inductive generalization. *Proceedings of the National Academy of Sciences*, 107(20), 9066–9071.
- Hegdedus, T. (1995). Generalized teaching dimensions and the query complexity of learning. *Proceedings of the eighth annual conference on Computational learning theory - COLT '95*. <https://doi.org/10.1145/225298.225311>
- Jara-Ettinger, J., Gweon, H., Schulz, L. E., & Tenenbaum, J. B. (2016). The naïve utility calculus: Computational principles underlying commonsense psychology. *Trends in Cognitive Sciences*, 20(8), 589–604. <https://doi.org/10.1016/j.tics.2016.05.011>
- Johnson, D. D. P., & Fowler, J. H. (2011). The evolution of overconfidence. *Nature*, 477, 317–320. <https://doi.org/10.1038/nature10384>
- Juni, M. Z. J., Gureckis, T. M., & Maloney, L. T. (2010). Integration of visual information across time. *Visual Science Society Meeting Abstract, August*.

- Kemp, C., & Tenenbaum, J. (2008). The discovery of structural form. *Proceedings of the National Academy of Sciences*, 105(31).
- Keysar, B., & Henly, A. S. (2002). Speakers' overestimation of their effectiveness [PMID: 12009039]. *Psychological Science*, 13(3), 207–212. <https://doi.org/10.1111/1467-9280.00439>
- Khan, F., Zhu, X., & Mutlu, B. (2011). How do humans teach: On curriculum learning and teaching dimension. *Advances in Neural Information Processing Systems (NIPS 24)*, 10687–10692.
- Kiani, R., Corthell, L., & Shadlen, M. N. (2014). Choice certainty is informed by both evidence and decision time. *Neuron*, 84(6), 1329–1342. <https://doi.org/10.1016/j.neuron.2014.12.015>
- Klemencic, M., Pupinis, M., & Kirdulyte, G. (2020). *Mapping and analysis of student centred learning and teaching practices : Usable knowledge to support more inclusive, high-quality higher education : Analytical report*. Publications Office of the European Commission; Directorate-General for Education, Youth, Sport; Culture. <https://doi.org/doi/10.2766/67668>
- Kline, M. A. (2015). How to learn about teaching: An evolutionary framework for the study of teaching behavior in humans and other animals. *The Behavioral and brain sciences*, 38. <https://doi.org/https://doi.org/10.1017/S0140525X14000090>
- Kruger, J., & Dunning, D. (1999). Unskilled and unaware of it: How difficulties in recognizing one's own incompetence lead to inflated self-assessments. *Journal of Personality and Social Psychology*, 77(6), 1121–1134. <https://doi.org/10.1037/0022-3514.77.6.1121>
- Kurt, V. (2011). The relative effectiveness of human tutoring, intelligent tutoring systems, and other tutoring systems. *Educational Psychologist*, 46(4), 197–221. <https://doi.org/10.1080/00461520.2011.611369>
- Kurvers, R. H., Hertz, U., Karpus, J., Balode, M. P., Jayles, B., Binmore, K., & Bahrami, B. (2021). Strategic disinformation outperforms honesty in competition for social influ-

- ence. *iScience*, 24(12), 103505. <https://doi.org/https://doi.org/10.1016/j.isci.2021.103505>
- Lake, B. M., Salakhutdinov, R., & Tenenbaum, J. B. (2015). Human-level concept learning through probabilistic program induction. *Science*, 350(6266), 1332–1338.
- Landrum, A. R., Cloudy, J., & Shafto, P. (2015). More than true: Developmental changes in the use of inductive strength for selective trust.
- Landrum, A. R., Eaves, J., B. S., & Shafto, P. (2015). Learning to trust and trusting to learn: A theoretical framework. *Trends in Cognitive Sciences*, 19(3), 109–111.
- Lee, E.-J. (2005). Effects of the influence agents sex and self-confidence on informational social influence in computer-mediated communication:: Quantitative versus verbal presentation. *Communication Research*, 32(1), 29–58. <https://doi.org/10.1177/0093650204271398>
- Leelawong, K., & Biswas, G. (2008). Designing learning by teaching agents: The betty's brain system. *I. J. Artificial Intelligence in Education*, 18, 181–208.
- Lieder, F., Griffiths, T., & Goodman, N. (2012). Burn-in, bias, and the rationality of anchoring. In F. Pereira, C. J. C. Burges, L. Bottou, & K. Q. Weinberger (Eds.), *Advances in neural information processing systems*. Curran Associates, Inc. <https://proceedings.neurips.cc/paper/2012/file/81e5f81db77c596492e6f1a5a792ed53-Paper.pdf>
- Lieder, F., Plunkett, D., Hamrick, J. B., Russell, S. J., Hay, N., & Griffiths, T. (2014). Algorithm selection by rational metareasoning as a model of human strategy selection. In *Advances in Neural Information Processing Systems 27*, edited by Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, 2870-78. Curran Associates, Inc., 27.
- Liu, W., Dai, B., Humayun, A., Tay, C., Yu, C., Smith, L. B., & Song, L. (2017). May 30). <http://arxiv.org/abs/1705.10470>
- Luce, R. D. (1959). *Individual choice behavior*. [Pages: xii, 153]. John Wiley.
- Maddox, W. T., & Ashby, F. G. (2004). Dissociating explicit and procedural-learning based systems of perceptual category learning. [Place: Netherlands Publisher: Elsevier Science]. *Behavioural Processes*, 66(3), 309–332. <https://doi.org/10.1016/j.beproc.2004.03.011>

- Mahmoodi, A., Bang, D., Olsen, K., Zhao, Y. A., Shi, Z., Broberg, K., Safavi, S., Han, S., Nili Ahmadabadi, M., Frith, C. D., Roepstorff, A., Rees, G., & Bahrami, B. (2015). Equality bias impairs collective decision-making across cultures. *Proceedings of the National Academy of Sciences*, 112(12), 3835–3840. <https://doi.org/10.1073/pnas.1421692112>
- 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.
- Martí, L., Mollica, F., Piantadosi, S., & Kidd, C. (2018). Certainty is primarily determined by past performance during concept learning. *Open Mind*, 2(2), 47–60. https://doi.org/10.1162/opmi_a.00017
- Mayer, R. E. (2004). Should there be a three-strikes rule against pure discovery learning? *American Psychologist*, 59(1), 14–19.
- McCandliss, B. D., Fiez, J. A., Protopapas, A., Conway, M., & McClelland, J. L. (2002). Success and failure in teaching the [r]-[l] contrast to Japanese adults: Tests of a Hebbian model of plasticity and stabilization in spoken language perception. *Cognitive, Affective & Behavioral Neuroscience*, 2(2), 89–108.
- Meng Yuan, T. L., Griffiths, & Fei, X. (2017). Inferring intentional agents from violations of randomness.
- Montessori, M., & Gutek, G. L. (2004). *The montessori method: The origins of an educational innovation: Including an abridged and annotated edition of maria montessori's the montessori method*. Lanham, Md: Rowman & Littlefield Publishers.
- Navarro, D. J., Dry, M. J., & Lee, M. D. (2012). Sampling assumptions in inductive generalization. *Cognitive Science*, 36(2), 187–223.
- Nosofsky, R. M. (1989). Further tests of an exemplar-similarity approach to relating identification and categorization. *Perception & psychophysics*, 45(4), 279–290. <https://doi.org/10.3758/bf03204942>

- Op de Beeck, H., Wagemans, J., & Vogels, R. (2001). Inferotemporal neurons represent low-dimensional configurations of parameterized shapes. *Nature Neuroscience*, 4(12), 1244–1252. <https://doi.org/10.1038/nn767>
- Paradise, R., & Rogoff, B. (2009). Side by side: Learning by observing and pitching in. *Ethos*, 37(1), 102–138. <https://doi.org/10.1111/j.1548-1352.2009.01033.x>
- Patel, D., Fleming, S. M., & Kilner, J. M. (2012). Inferring subjective states through the observation of actions. *Proceedings of the Royal Society B: Biological Sciences*, 279(1748), 4853–4860. <https://doi.org/10.1098/rspb.2012.1847>
- Patil, K. R., Zhu, X., Kopec, L., & Love, B. C. (2014). Optimal teaching for limited-capacity human learners., 2465–2473.
- Pescetelli, N., Rees, G., & Bahrami, B. (2016). The perceptual and social components of metacognition. *Journal of Experimental Psychology: General*, 145(8), 949–965. <https://doi.org/10.1037/xge0000180>
- Pouget, A., Drugowitsch, J., & Kepecs, A. (2016). Confidence and certainty: Distinct probabilistic quantities for different goals. *Nature Neuroscience*, 19(3), 366–374. <https://doi.org/10.1038/nn.4240>
- Price, P. C., & Stone, E. R. (2004). Intuitive evaluation of likelihood judgment producers: Evidence for a confidence heuristic. *Journal of Behavioral Decision Making*, 17(1), 39–57. <https://doi.org/10.1002/bdm.460>
- Pulford, B. D., Colman, A. M., Buabang, E. K., & Krockow, E. M. (2018). The persuasive power of knowledge: Testing the confidence heuristic. *Journal of Experimental Psychology: General*, 147(10), 1431–1444. <https://doi.org/10.1037/xge0000471>
- Rafferty, A. N., LaMar, M. M., & Griffiths, T. L. (2015). Inferring learners' knowledge from their actions. *Cognitive Science*, 39(3), 584–618.
- Rafferty, A. N., Brunskill, E., Griffiths, T. L., & Shafto, P. (2011). Faster teaching by pomdp planning (G. Biswas, S. Bull, J. Kay, & A. Mitrovic, Eds.), 280–287.
- Rahnev, D., Desender, K., Lee, A. L. F., Adler, W. T., Aguilar-Lleyda, D., Akdogan, B., Arbuzova, P., Atlas, L. Y., Balci, F., Bang, J. W., Bague, I., Birney, D. P., Brady, T. F.,

- Calder-Travis, J., Chetverikov, A., Clark, T. K., Davranche, K., Denison, R. N., Dildine, T. C., ... Zylberberg, A. (2020). The Confidence Database. *Nature Human Behaviour*, 4(3), 317–325. <https://doi.org/10.1038/s41562-019-0813-1>
- Rhodes, M., Bonawitz, E., Shafto, P., Chen, A., & Caglar, L. (2015). Controlling the message: Preschoolers' use of information to teach and deceive others. *Frontiers in Psychology*, 6(867).
- Rhodes, M., Brickman, D., & Gelman, S. A. (2008). Sample diversity and premise typicality in inductive reasoning: Evidence for developmental change. *Cognition*, 108(2), 543–556.
- 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. <https://doi.org/https://doi.org/10.1037/xhp0000940>
- Rothe, A., Lake, B. M., & Gureckis, T. M. (2018). Do People Ask Good Questions? *Computational Brain & Behavior*, 1(1), 69–89. <https://doi.org/10.1007/s42113-018-0005-5>
- Rouder, J. N., Speckman, P. L., Sun, D., Morey, R. D., & Iverson, G. (2009). Bayesian t tests for accepting and rejecting the null hypothesis. *Psychonomic Bulletin & Review*, 16(2), 225–237.
- Rubio-Fernandez, P. (2019). Overinformative speakers are cooperative: Revisiting the gricean maxim of quantity. *Cognitive Science*, 43(11), e12797. <https://doi.org/https://doi.org/10.1111/cogs.12797>
- Sah, S., Moore, D. A., & MacCoun, R. J. (2013). Cheap talk and credibility: The consequences of confidence and accuracy on advisor credibility and persuasiveness. *Organizational Behavior and Human Decision Processes*, 121(2), 246–255. <https://doi.org/10.1016/j.obhdp.2013.02.001>

- Salmon, J. P., McMullen, P. A., & Filliter, J. H. (2010). Norms for two types of manipulability (graspability and functional usage), familiarity, and age of acquisition for 320 photographs of objects. *Behavior Research Methods*, 42(1), 82–95.
- Salvatier, J., Wiecki, T. V., & Fonnesbeck, C. (2016). Probabilistic programming in python using pymc3. *PeerJ Computer Science*, 2, e55.
- Sanborn, A., & Griffiths, T. (2007). Markov chain monte carlo with people (J. Platt, D. Koller, Y. Singer, & S. Roweis, Eds.). 20. <https://proceedings.neurips.cc/paper/2007/file/89d4402dc03d3b7318bbac10203034ab-Paper.pdf>
- Sanborn, A. N., & Chater, N. (2016). Bayesian brains without probabilities. *Trends in Cognitive Sciences*, 20(12), 883–893. <https://doi.org/https://doi.org/10.1016/j.tics.2016.10.003>
- Sanders, J. I., Hangya, B., & Kepecs, A. (2016). Signatures of a statistical computation in the human sense of confidence. *Neuron*, 90(3), 499–506. <https://doi.org/10.1016/j.neuron.2016.03.025>
- Searcy, S. R., & Shafto, P. (2016). Cooperative inference: Features, objects, and collections. *Psychological Review*, 123(5), 510–533.
- Shafto, P., & Goodman, N. D. (2008). Teaching games: Statistical sampling assumptions for pedagogical situations. *Proceedings of the 30th Annual Conference of the Cognitive Science Society*.
- Shafto, P., Goodman, N. D., & Frank, M. C. (2012). Learning from others. *Perspectives on Psychological Science: A Journal of the Association for Psychological Science*, 7(4), 341–351.
- Shafto, P., Eaves, B., Navarro, D. J., & Perfors, A. (2012). Epistemic trust: Modeling children's reasoning about others' knowledge and intent. *Developmental science*, 5(3), 436–447.
- 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.
- Simard, P. Y., Amershi, S., Chickering, D. M., Pelton, A. E., Ghorashi, S., Meek, C., Ramos, G., Suh, J., Verwey, J., Wang, M., & Wernsing, J. (2017). Machine teaching: A new

- paradigm for building machine learning systems. <https://doi.org/10.48550/ARXIV.1707.06742>
- Smith, S. G., & Sherwood, B. A. (1976). Educational uses of the plato computer system. *Science*, 192(4237), 344–352. <http://www.jstor.org/stable/1742096>
- Song, C., Kanai, R., Fleming, S. M., Weil, R. S., Schwarzkopf, D. S., & Rees, G. (2011). Relating inter-individual differences in metacognitive performance on different perceptual tasks. *Consciousness and Cognition*, 20(4), 1787–1792. <https://doi.org/10.1016/j.concog.2010.12.011>
- Sorkin, R. D., Hays, C. J., & West, R. (2001). Signal-detection analysis of group decision making. *Psychological Review*, 108(1), 183–203. <https://doi.org/10.1037/0033-295X.108.1.183>
- Sperber, D., Clament, F., Heintz, C., Mascaro, O., Mercier, H., Origgi, G., & Wilson, D. (2010). Epistemic vigilance. *Mind & Language*, 25(4), 359–393. [https://doi.org/https://doi.org/10.1111/j.1468-0017.2010.01394.x](https://doi.org/10.1111/j.1468-0017.2010.01394.x)
- Team, E. D. (2016). Stan modeling language users guide and reference manual, version 2.15.0. <http://mc-stan.org>
- Tenenbaum, J. B., & Griffiths, T. L. (2001). Generalization, similarity and Bayesian inference. [Place: United Kingdom Publisher: Cambridge University Press]. *Behavioral and Brain Sciences*, 24(4), 629–640. <https://doi.org/10.1017/S0140525X01000061>
- Tenney, E. R., MacCoun, R. J., Spellman, B. A., & Hastie, R. (2007). Calibration trumps confidence as a basis for witness credibility. *Psychological Science*, 18(1), 46–50. <https://doi.org/10.1111/j.1467-9280.2007.01847.x>
- Tenney, E. R., Meikle, N. L., Hunsaker, D., Moore, D. A., & Anderson, C. (2019). Is overconfidence a social liability? the effect of verbal versus nonverbal expressions of confidence. *Journal of Personality and Social Psychology*, 116(3), 396–415. <https://doi.org/10.1037/pspi0000150>
- Tenney, E. R., Small, J. E., Kondrad, R. L., Jaswal, V. K., & Spellman, B. A. (2011). Accuracy, confidence, and calibration: How young children and adults assess credibility.

- [Place: US Publisher: American Psychological Association]. *Developmental Psychology*, 47(4), 1065–1077. <https://doi.org/10.1037/a0023273>
- Tenney, E. R., Spellman, B. A., & MacCoun, R. J. (2008). The benefits of knowing what you know (and what you don't): How calibration affects credibility. *Journal of Experimental Social Psychology*, 44(5), 1368–1375. <https://doi.org/10.1016/j.jesp.2008.04.006>
- Thomas, J. P., & McFadyen, R. G. (1995). The confidence heuristic: A game-theoretic analysis. *Journal of Economic Psychology*, 16(1), 97–113. [https://doi.org/10.1016/0167-4870\(94\)00032-6](https://doi.org/10.1016/0167-4870(94)00032-6)
- Thornton, A., & Raihani, N. J. (2008). The evolution of teaching. *Animal Behaviour*, 75(6), 1823–1836. <https://doi.org/10.1016/j.anbehav.2007.12.014>
- Tomasello, M., Kruger, A. C., & Ratner, H. H. (1993). Cultural learning. *Behavioral and Brain Sciences*, 16(3), 495–511. <https://doi.org/10.1017/S0140525X0003123X>
- Tversky, A., & Kahneman, D. (1974). Judgement under uncertainty: Heuristics and biases. *Science*, 185, 1124–1131. <https://doi.org/10.1126/science.185.4157.1124>
- VanLehn, K., Graesser, A. C., Jackson, G. T., Jordan, P., Olney, A., & Rosé, C. P. (2007). When are tutorial dialogues more effective than reading? *Cognitive Science*, 31(1), 3–62. <https://doi.org/10.1080/03640210709336984>
- Vul, E., Goodman, N., Griffiths, T. L., & Tenenbaum, J. B. (2014). One and done? optimal decisions from very few samples. *Cognitive Science*, 38(4), 599–637. <https://doi.org/10.1111/cogs.12101>
- Wallander, D., & Boynton. (2013). Systematic overweighting of early items in serial cue integration. *Visual Cognition*, 21:6, 689–692.
- Wang, F. (2014). Pomdp framework for building an intelligent tutoring system, 233–240. <https://doi.org/10.5220/0004801702330240>
- Wang, P., Wang, J., Paranamana, P., & Shafto, P. (2020). A mathematical theory of cooperative communication (H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, & H. Lin, Eds.). 33, 17582–17593.

- Warner, R., Stoess, T., & Shafto, P. (2011). Reasoning in teaching and misleading situations. *Proceedings of the 33rd Annual Conference of the Cognitive Science Society*.
- Weisstein, E. W. (n.d.). Square triangle picking. <https://mathworld.wolfram.com/SquareTrianglePicking.html>
- Wood, D., Bruner, J. S., & Ross, G. (1976). The role of tutoring in problem solving*. *Journal of Child Psychology and Psychiatry*, 17(2), 89–100. <https://doi.org/10.1111/j.1469-7610.1976.tb00381.x>
- Xiong, C., Van Weelden, L., & Franconeri, S. (2020). The curse of knowledge in visual data communication. *IEEE Transactions on Visualization and Computer Graphics*, 26(10), 3051–3062. <https://doi.org/10.1109/TVCG.2019.2917689>
- Xu, F., & Tenenbaum, J. (2007). Sensitivity to sampling in Bayesian word learning. *Developmental Science*, 10(3), 288–297.
- Yang, S. C.-H., & Shafto, P. (2017a). Teaching versus active learning: A computational analysis of conditions that affect learning. *Proceedings of the 39th Annual Conference of the Cognitive Science Society*, (4), 3560–3565.
- Yang, S. C.-H., & Shafto, P. (2017b). Teaching versus active learning: A computational analysis of conditions that affect learning.
- Yang, S. C.-H., Vong, W. K., Yu, Y., & Shafto, P. (2019). A unifying computational framework for teaching and active learning. *Topics in Cognitive Science*.
- Yang, S. C.-H., Wolpert, D. M., & Lengyel, M. (2018). Theoretical perspectives on active sensing. *Current Opinions in Behavioural Science*, 11, 100–108.
- Yang, S. C.-H., Yu, Y., Givchi, A., Wang, P., Vong, W. K., & Shafto, P. (2018). Optimal cooperative inference. *arXiv:1705*.
- Yates, J., Price, P. C., Lee, J.-W., & Ramirez, J. (1996). Good probabilistic forecasters: The ‘consumer’s’ perspective. *International Journal of Forecasting*, 12(1), 41–56. [https://doi.org/10.1016/0169-2070\(95\)00636-2](https://doi.org/10.1016/0169-2070(95)00636-2)
- Yurovsky, D., Smith, L. B., & Yu, C. (2013). Statistical word learning at scale: The baby’s view is better. <https://doi.org/10.1111/desc.12036>

- Zhu. (2013). Machine teaching for bayesian learners in the exponential family. *Proceedings of the 26th International Conference on Neural Information Processing Systems*, 2, 1905–1913.
- Zhu, J., Sanborn, A., & Chater, N. (2018). Mental sampling in multimodal representations (S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, & R. Garnett, Eds.). 31. <https://proceedings.neurips.cc/paper/2018/file/b4a721cfb62f5d19ec61575114d8a2d1-Paper.pdf>
- Zilles, S., Lange, S., Holte, R., & Zinkevich, M. (2011). Models of cooperative teaching and learning. *Journal of Machine Learning Research*, 12, 349–384.