Modeling How Humans Judge Dot-Label Relations in Point Cloud Visualizations – Supplemental Material

Martin Reckziegel, Linda Pfeiffer, Christian Heine, and Stefan Jänicke

Abstract— This supplemental material to the paper entitled 'Modeling How Humans Judge Dot-Label Relations in Point Cloud Visualizations' contains a discussion on how we optimized our experimental design for model discrimination, additional diagrams, and data tables to support our analysis and findings, as well as descriptions of supplemental files and further observations.

1 Optimal Stimulus Design

In the main paper we described twelve model classes that predict how humans relate textual labels to dots in graphical representations comprising them. We would like to know which one of them gives the most accurate predictions for real world instances. In such instances, there are usually a huge number of dots and labels. It is infeasible to run an experiment in which every potential factor influencing the outcome is varied systematically, because of the combinatorial explosion of the number of stimuli required. Plus, to discriminate between the classes of models we hypothesize, that it is sufficient to only test the most 'telling' stimuli, i.e., the stimuli for which the different model classes perform very differently so that our experiment produces the most amount of information about which model is the most accurate using the least amount of trials. Intuitively, one need not check instances for which all models predict the same or very similar results. For probabilistic models such as ours, characterizing similarity in results is non-trivial. In such situations, one can use design optimization [3, 5].

Design optimization is a statistical technique to select stimuli such that an experiment will discriminate the models best. Since design optimization is little known and used in visualization research, we would like to give a short introduction. As the topic is rather general and mathematically involved, we put it in this supplemental material to distract as little as possible from the main 'narrative'. Note however, that the design optimization is a dispensable part of our experimental design, i.e. one could repeat our experiment without the optimization, but at the cost of using more trials to obtain similar results.

1.1 Information-Theoretic Framework

This section briefly outlines an information-theoretic framework for design optimization. Imagine a set of generators, each of which gives, when presented with an experimental design d from the set of possible designs (e.g. sequences of stimuli), a response r. The response r is drawn from a probability distribution specific to each generator. Imagine now, that I pick a generator g, but do not tell you which one. However, I leave you with the choice of the design d and will use the generator g to give you a response r for it (this corresponds to running an experiment). From r you can make an informed guess about g. To guess g well, you should pick d in a way that r tells you the most about the unknown g.

This 'game' can be modeled with probability theory and information theory as follows: Letting G denote the random variable over the space of generators and R the random variable over the space of responses, we can describe the probabilistic process with the joint probability mass function $P(G=g,R=r|D=d)$ conditional on D, the set of possible designs. Note that while the marginal distribution for R, $P(R=r|D=d)$, depends on the design d, the marginal distribution for G, $P(G=g)$, does not, because neither do I select g based on your d, nor can you select d based on the g unknown to you.

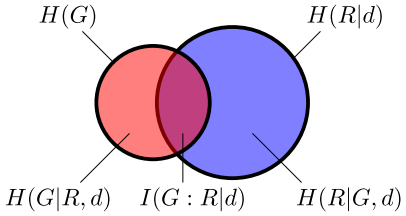


Fig. 1: Information-theoretic view on experiments. A design d will have experimental results R from which one can reduce the uncertainty about an unknown G. See main text for description of the quantities.

Using these probability functions, we can derive information-theoretic quantities [2] to model the amount of information about G and the amount of information present in R. Generally, the entropy $H(X)$ can be seen as a measure for the uncertainty about or information present in a random variable X. It can be computed as the expected value of the quantity $-\log P(X)$. The joint entropy of two random variables $H(X,Y)$ is a measure for the uncertainty about or information present in two random variables. It can be computed as the expected value of the quantity $-\log P(X,Y)$ and is less than or equal to the sum of the entropy for each variable. The mutual information $I(X:Y)$ between two random variables X and Y measures the amount of information shared between the two variables, and is equal to $H(X)+H(Y)-H(X,Y)$. One can adapt these measures for conditional probabilities, e.g. $H(X|Y)$ is the expected value of $-\log P(X|Y)$. All these measures are non-negative. We can assign an interpretation to these measures as follows (see also Figure 1):

$H(G)$ – the uncertainty about G

$H(R|d)$ – the amount information in the results R when using experimental design d

$I(G:R|d)$ – the reduction in uncertainty about G considering the results R from an experiment with design d

$H(G|R,d)$ – the uncertainty about G left after considering the results R from an experiment with design d

$H(R|G,d)$ – the information in the experimental results R that tell nothing about G, i.e. useless information.

There are now multiple strategies for selecting d. The first strategy is to select d which minimizes $H(G|R,d)$, i.e. which reduces the uncertainty about G the most, or equivalently that maximizes $I(G:R|d)$, i.e. which maximizes the information about G in R ([1, 4]). Using elementary transformations, one can show that this is equal to maximizing the expected value of the log posterior probability $P(G|R,d)$. This method

- *M. Reckziegel and C. Heine are with Leipzig University. E-mail: {reckziegel,heine}@informatik.uni-leipzig.de.*
- *Linda Pfeiffer is with German Aerospace Center DLR. E-mail: linda.pfeiffer@dlr.de.*
- *Stefan Jänicke is with University of Southern Denmark. E-mail: stjaenicke@imada.sdu.dk.*

has the disadvantage that it ignores $H(R|G,d)$, i.e. the amount of useless information produced in the experiment. Therefore, this strategy does not penalize long and complicated experiments, which we wish to avoid. Since the amount of uncertainty about G can never be reduced fully when the generators are probabilistic, the optimal design d would comprise infinite stimuli – an infeasible option. The second strategy accounts for useless information and minimizes $H(G|R,d)+H(R|G,d)$ instead. This measure equals $H(G,R|d)-I(G:R|d)$. Since the joint entropy typically grows faster than the mutual information, the optimal design therefore comprises zero stimuli. Thus, we used a third strategy which maximizes what we will call the *design efficiency*:

$$\arg\max_d \frac{I(G:R|d)}{H(G,R|d)}$$

This proportion of shared vs. total information has the advantage of being unitless and lies in $[0,1]$. If the efficiency of the design is 0, the experiment tells nothing about G and if it is 1 then G can be fully known and no useless information is produced. $I(G:R|d)$ is bounded from above by $H(G)$, which does not depend on the number of stimuli in the design, and $H(G,R|d)$ is bounded by $H(G)+H(R|d)$. Therefore, increasing the number of stimuli in the design will have diminishing returns, i.e., its efficiency will increase up to an optimal number of stimuli and decrease afterwards. Informal experiments within our problem setting showed that optimizing this measure also gives good results concerning the reduction of our uncertainty about G: the best design with regard to the former measure was only 7% worse than the best design with regard to the latter, but significantly better in reducing useless information.

1.2 Application to our Experimental Design

The above thought experiment is applied to our problem setting as follows: The generators correspond to the models described in the main paper. Stimuli are pairs of images and query points. We presume that one of our models predicts the human's responses well, although we do not yet know which one.

There are some challenges when applying the above to our experiment which we will discuss in this section. First, the models do not consider memory effects humans are subject to. Second, the model classes have real-valued parameters, i.e., we actually have an infinite number of models. Third, finding the optimal design is intractable in our case, because the time needed to compute the design efficiency grows exponential with the number of stimuli. Finally, picking the design that discriminates models best may produce stimuli that are not representative of the target application.

The first challenge involves an idealizing assumption. Our pilot study indicated that memory effects can occur with humans, i.e., if two stimuli are similar and presented with only few other stimuli in between, humans provide consistent answers and reported this in the questionnaire. Our models do not contain this memory effect, i.e., if the design of an experiment is a sequence of stimuli $d=((I^{(1)},\mathbf{x}^{(1)}),\dots,(I^{(n)},\mathbf{x}^{(n)}))$ and the response of the experiment is a list of labels $r=(l^{(1)},\dots,l^{(n)})$ then the likelihood

$$P(R=r|G=M,d)=\prod_{i=1}^{n} p^{M}_{I^{(i)}}(l^{(i)},\mathbf{x}^{(i)}) \tag{1}$$

is independent of the order of stimuli for each model M. We controlled for this effect by generating many different stimuli and presenting them in a random order for each participant of the study. Nevertheless, our models are only valid for situations without memory effects.

The second challenge results from the real-valued parameters that our models have. When computing the values of the information-theoretic measures above, one would need to integrate over functions for which we know no analytic solution. Integrating numerically is too expensive computationally. However, most of the parameters can be bounded, because extreme parameter values are highly implausible. Therefore we selected a finite number of models that mark the vertices of a region in parameter space bounding the set of plausible values. We also added models, whose parameter values were fitted for the results of our pilot study. This finite set of models was exclusively used to generate 'the most informative points', they have no impact in the resulting analysis. Our prior probability for each model is the same.

The third challenge is the intractability of the above method. Computing the information-theoretic measures above involves computing multiple nested sums that range over the labels in each stimulus. Let n be the number of stimuli, m the number of models, k the maximum number of labels in each image, then the run time is bounded from above by $O(k^n m)$. Therefore we split the design in chunks of 4 stimuli and optimize each chunk at a time independently of the other chunks.

Since there is an infinite number of possible designs, one has to describe the designs as a set of real-valued stimuli parameters. Then, the optimal design can, in principle, be found by classical optimization. However, this may result in stimuli that may be well suited to discriminate the models, but are no longer representative for the target application. Therefore, we use stimuli generators for all image properties and only optimize the position $\mathbf{x}$ of the query dot. Since derivatives for the design efficiency measure are difficult to compute, in particular because of the force-directed deformation of space, we use a slightly modified greedy approach. We successively add new stimuli to an initially empty design chunk, each time choosing $\mathbf{x}$ optimally to supplement the points already in the design chunk, until the chunk reaches size 4. We pick $\mathbf{x}$ from a large set of points which were randomly placed in an image maintaining a minimum distance. Then, for each point we cycle again through all valid locations, replacing the existing positions with new ones if they increase design efficiency until no change occurs for a full loop over the chunk.

The final challenge is that this method tests only query points that discriminate the given models best. Untested points are uncritical as they will affect the likelihood scores of the models in the same way so that there is no additional difference in likelihood scores (otherwise they would have been good candidates for testing and would have been picked). But one should not use additional model classes after selecting the query points because this can result in models that perform better than the original models, but only because there are no 'critical stimuli' tested, i.e. stimuli for which the new models would perform badly.

1.3 Implementation

Computing the design efficiency is straight-forward. We use the following identities that can be shown elementary to simplify the formula:

$$\frac{I(G:R|d)}{H(G,R|d)}=\frac{H(G)+H(R|d)}{H(G,R|d)}-1$$

with

$$H(G)=-\sum_g P(G=g)\log P(G=g)$$

$$H(R|d)=-\sum_r P(R=r|d)\log P(R=r|d)$$

$$H(G,R|d)=-\sum_r P(G=g,R=r|d)\log P(G=g,R=r|d)$$

and

$$P(R=r|d)=\sum_g P(G=g,R=r|d)\log P(G=g,R=r|d)$$

$$P(G=g,R=r|d)=P(G=g)P(R=r|G=g,d).$$

1.4 Evaluation

Design optimization is preferable to choosing query points randomly or by hand, since the space of the designs in which the optimization takes place also contains the random or manual choice. Unless the optimization problem cannot be formulated or solved numerically, manual choice cannot perform better than a search through all the options. We performed a small evaluation to assess the improvement of using design optimization. For this, we compared the efficiency for images generated for our experiment using (a) design optimization, (b)

Table 1: Results for the two-sided pair-wise sign tests for the differences in AIC scores between any two of the twelve model classes. The tests were performed for a significance level of $\alpha = 0.05$, using Bonferroni correction. For cells with a red background the model class of the corresponding column performs better than the model class of the corresponding row and vice versa for cells with a blue background. For cells with a white background no significant difference between the row and column model classes was found.

VLC	38.4	46.2	46.6	19.8	36.0	39.2	39.8	31.9	33.1	37.2	37.7
-38.4	VLB	29.1	34.8	-51.0	1.1	20.7	22.5	-26.8	24.4	18.6	27.2
-46.2	-29.1	VCB	34.8	-51.7	-28.7	-5.5	10.3	-22.4	14.8	21.3	23.7
-46.6	-34.8	-34.8	VCS	-54.9	-31.0	-17.1	-10.8	-29.2	13.7	19.3	23.5
-19.8	51.1	51.7	54.9	ALC	51.1	46.2	50.5	31.7	42.2	41.2	44.1
-36.0	-1.1	28.7	31.0	-51.2	ALB	24.8	26.6	-23.3	22.9	18.2	23.7
-39.2	-20.7	5.5	17.1	-46.2	-24.8	ACB	22.4	-17.8	19.0	21.9	24.5
-39.8	-22.5	-10.3	10.8	-50.5	-26.6	-22.4	ACS	-23.0	18.4	18.5	21.6
-31.9	26.8	22.4	29.2	-31.7	23.3	17.8	23.0	ELC	23.0	20.6	26.7
-33.1	-24.4	-14.8	-13.7	-42.2	-22.9	-21.9	-18.4	-23.0	ELB	11.0	20.6
37.2	18.6	21.3	19.3	41.2	18.2	21.9	18.5	20.6	11.0	ECB	36.9
37.7	-27.2	23.7	-23.5	-44.1	-23.7	-24.5	-21.6	-26.7	-20.6	-36.9	ECB

random selection, (c) manual selection. Since we used manual selection for our pilot study and implemented design optimization only for the full study the comparison is most likely unbiased. For 4 query points in the same image, we observed an average efficiency of 0.242 in the design optimization condition and an average efficiency of 0.129 in the random condition. The best random selection out of 10 had an efficiency of 0.168. We furthermore observed that manual selection had an average efficiency around 0.160, i.e., better than average, but worse than best out of 10 random draws. We estimate that without design optimization one would need approximately 50 percent more stimuli to obtain similar results.

2 Stimulus Data

The file 'stimuli.json' contains all the information to reproduce the stimuli in machine readable form. The stimuli are listed ordered by their index in the array 'stimuli'. Each item of the array contains the dimensions of the stimulus in 'pixel' units and the number of centimeters per 'pixel', specifying the length of each unit after calibration. Furthermore, the array 'labels' specifies the font size, label text, position, and bounding box of all labels as well as the array 'points' specifies the center positions of all dots of the stimulus. The property 'stimulusPoint' specifies the center position of the query dot. The images of all 214 stimuli can be found as PNG files in the folder named stimuli. The files are named by the stimulus index. The 'Arimo' font[1] is used to render the labels and each dot has a radius of 3 pixel.

3 Sample Experiment

The video file 'sample-experiment-setting.mp4' shows a screencast of the complete procedure of the online experiment. For the purpose of demonstration, the size of each of the four blocks was reduced to 8 stimuli per block.

4 Participant Results Data

The result data of the participants is given as CSV files. The first row is the header listing the column names. The cells are not escaped.

[1] https://www.fontsquirrel.com/fonts/arimo

4.1 Response Times

The response time equals the time between the start of the visibility of a stimulus and the time the label was clicked by the user in milliseconds. These can be found in the file 'blocktimedata.csv'. The first column lists the index of the stimulus, the next 94 columns the response times for each participant, labeled by their id. The last two columns list the mean and standard deviation of the response times per stimulus.

4.2 Participant Answers

Similarly, the actual answers of the participants can be found in the file 'blocklabeldata.csv'. The first column lists the index of the stimulus, the next 94 columns the index of the clicked label of each participant, labeled by their id. The last column lists the entropy calculated from the histogram of the different clicked labels for each stimulus.

4.3 Outlier

Both CSV files contain all participants' data including the outliers. As stated in the main paper, we removed two participants (p46, p81) based on the MDS scatter plots of the response times and two participants (p12, p53) based on the MDS scatter plots of the participant answers. In addition, we found participants p77 and p78 to have identical data with respect to both, response times and answers, such that we removed one of them as a duplicate submission.

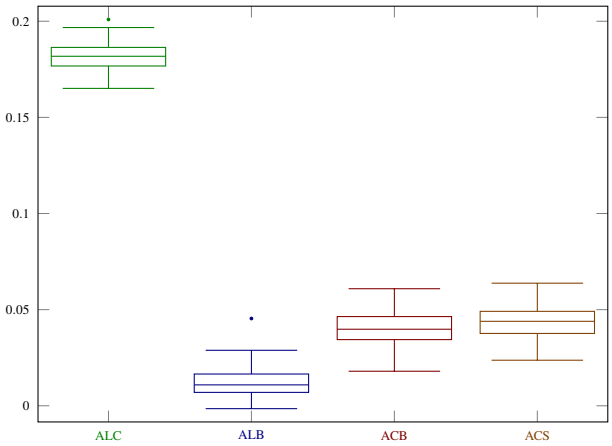
5 Analysis Data

5.1 Pair Tests

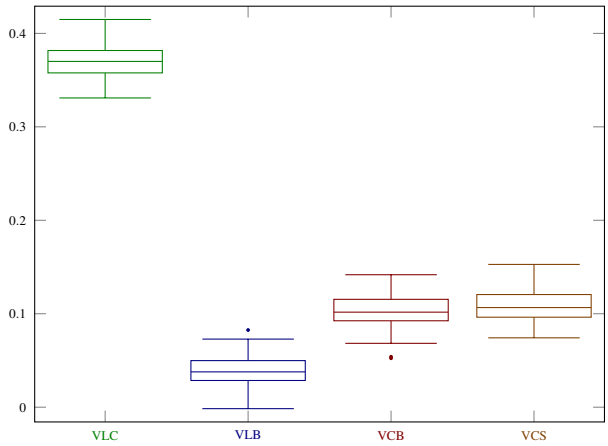
Table 1 shows the result of the two-sided pair-wise sign tests for the differences in AIC scores between any two of the twelve model classes. Counting the cells with a blue background for each row, we obtain the number of model classes against the class of the respective row performs better. Except for the VLB vs. ALB pair which does not show significant differences, we can order the model classes descending by performance, such that ECS $>$ ECB $>$ ELB $>$ VCS $>$ ACS $>$ VCB $>$ ACB $>$ VLB, ALB $>$ ELC $>$ ALC $>$ VLC.

5.2 Bootstrapped Parameters

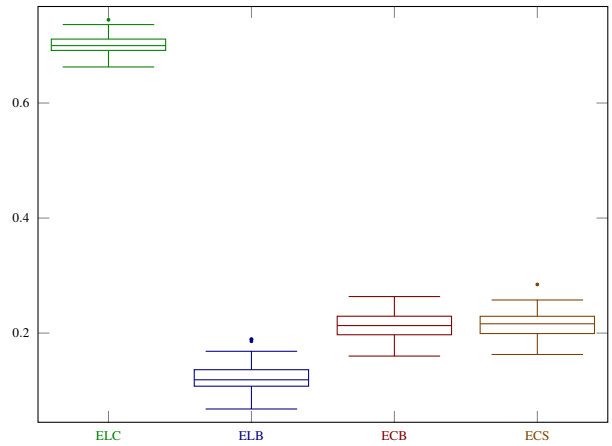
Fig. 2 shows box plots for each parameter of each model class resulting out of the bootstrapped maximum likelihood estimation. We generated 100 bootstrap samples, each generated as follows. We picked one of



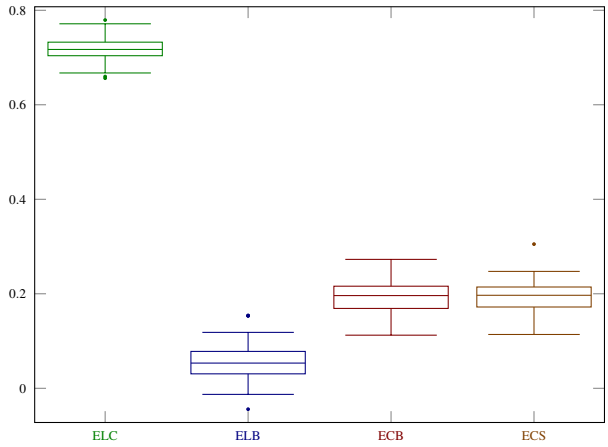
(a) Parameter a (area) of the area-weighted model class **A**



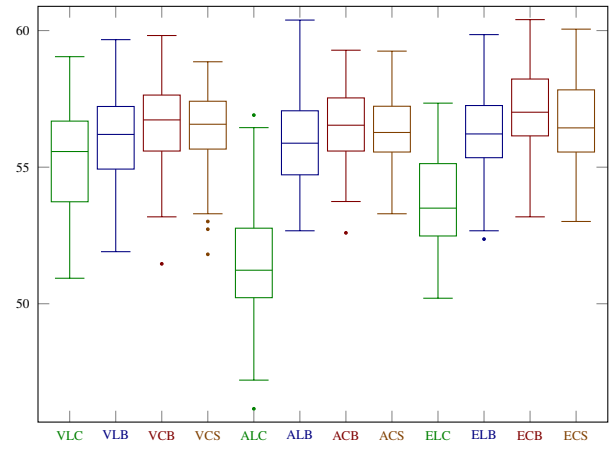
(b) Parameter h (height) of the vertically-weighted model class **V**



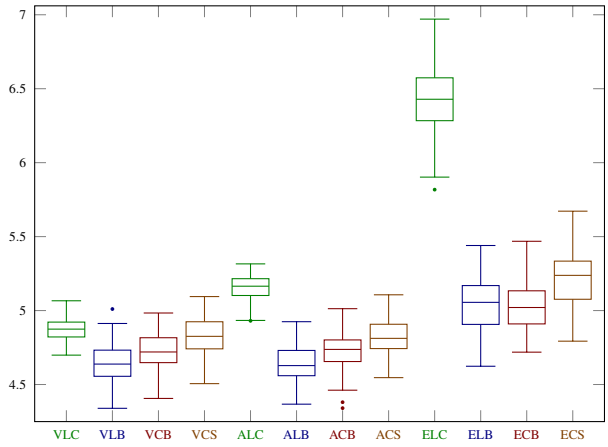
(c) Parameter w (width) of the elliptically-weighted model class **E**



(d) Parameter h (height) of the elliptically-weighted model class **E**



(e) Parameter f (force) of each model classes



(f) Parameter c (certainty) of each model class

Fig. 2: Results of the bootstrapped maximum likelihood estimation for the parameters of the model classes. Each box plot represents the parameter fittings based on 100 bootstrap samples.

the participants out of the complete set of participants (89) randomly with uniform distribution and added her answers to the sample set. We repeated this step 89 times to generate 89 'virtual' participants by drawing with replacement.

5.3 Likelihood Ratio Tests

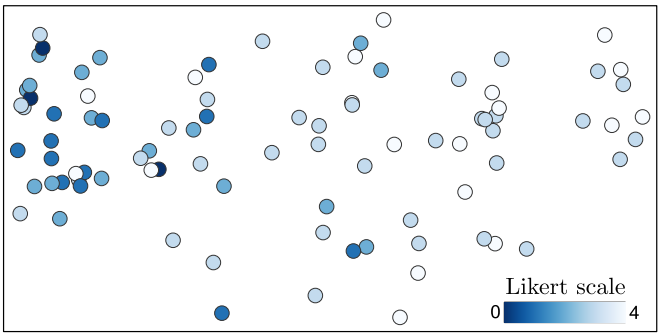
We computed for each model class, defined by its distance type and font size weighting type, submodel classes, wherein the influence of the parameters representing a specific judgment criterion is disabled. We fitted the parameters and applied a likelihood-ratio test to each submodel and the corresponding complete model. The results are shown in Table 2. For each model class the complete models performed significantly better than their submodels. As such, all parameters and the related judgment criteria significantly contribute to the overall model performance.

5.4 Further Observations

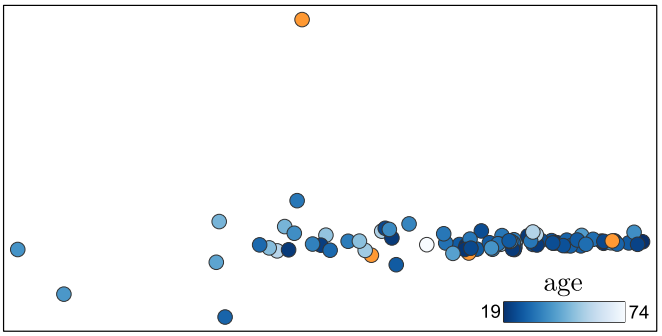
As described in the main paper to identify outlier, we created scatter plots using classical multidimensional scaling (MDS), minimizing the squared error between the Euclidean distances of the coordinates and two dissimilarity scores. The resulting plots (with outliers removed) shown in Fig. 3 reveal two interesting patterns. Fig. 3a shows the MDS plot computed from pairwise dissimilarities between the participants based on the number of times they disagree in their answers. Coloring the plot according to the Likert-scale ratings for the impact of presence of dots shows a rising impact along the left-right axis for the participant answers. This suggests that people whose decision was highly influenced by the presence of other dots do considerably differ in their answers from people that do not care about the presence of other dots. Fig. 3b shows the MDS plot computed from the participants dissimilarities based on the Euclidean distance between their vectors which components are yielded by their answer times of each stimulus. The participants' positions here align along one direction. This effect indicates that there is one variable that explains most differences in timings. Coloring the plot according to age shows that along this direction ages decrease, suggesting that this variable may explain these differences. Also, Kendall's tau showed a weak but significant correlation between age and the mean response times per person with medium/small effect ($\tau = .2480, p = .0010$). This is consistent with findings for general perceptual choice experiments.

Furthermore, we hypothesized that the diversity in the participants answers per stimulus would be reflected by the answer times. We used the entropy of the histogram of the different labels chosen as measure for the diversity of a stimulus. Our hypothesis was confirmed by a positive Pearson correlation ($\tau = .6853, p < .0001$) calculated for this measure against the mean answer time of each stimulus. This is consistent with findings for general perceptual choice experiments.

Apart from the mentioned effects we did not find further patterns. Nor did the tests for association (depending on the scales of the variables we used Kendall's rank correlation or coefficient η) between socio-demographic data and Likert scale ratings show any medium sized or strong correlations.



(a) The distance between two participants is the number of different label choices across all stimuli. The coloring is based on the Likert scale ratings for the impact by presence of dots. The higher the rating the brighter the color.



(b) The distance between two participants is the euclidean distance between their vectors which components are yielded by their answer times of each stimulus. The coloring is based on the age of the participants. The higher the age the brighter the color. An orange color indicates the absence of the participant's answer to the age.

Fig. 3: MDS scatter plots for the distances between the participants based on two dissimilarities.

References

[1] D. R. Cavagnaro, J. I. Myung, M. A. Pitt, and J. V. Kujala. Adaptive design optimization: A mutual information-based approach to model discrimination in cognitive science. *Neural computation*, 22(4):887–905, 2010.

[2] T. M. Cover and J. A. Thomas. *Elements of Information Theory*. Wiley-Interscience, 2nd ed., 2006.

[3] J. I. Myung and M. A. Pitt. Optimal experimental design for model discrimination. *Psychological review*, 116(3):499, 2009.

[4] L. Paninski. Asymptotic theory of information-theoretic experimental design. *Neural Computation*, 17(7):1480–1507, 2005.

[5] E. G. Ryan, C. C. Drovandi, J. M. McGree, and A. N. Pettitt. A review of modern computational algorithms for bayesian optimal design. *International Statistical Review*, 84(1):128–154, 2016.

Table 2: Results of the likelihood ratio tests relating submodels, reduced by single parameters respectively judgment criteria, to the corresponding complete models

model	**distance type**	**disabled parameter**	**likelihood ratio**	**critical χ^2 value**	**df**	α	**p-value**	**Bonferroni corrected p-value**
A	CB	c	33324.5556	3.84145882	1	0.05	0	0
A	CB	f	1625.9119	3.84145882	1	0.05	0	0
A	CB	a	163.2827	3.84145882	1	0.05	2.17E-25	7.81208E-24
A	CS	c	33734.7131	3.84145882	1	0.05	0	0
A	CS	f	1587.5170	3.84145882	1	0.05	0	0
A	CS	a	192.7879	3.84145882	1	0.05	7.83E-44	2.81919E-42
A	LB	c	32893.8679	3.84145882	1	0.05	0	0
A	LB	f	1524.6170	3.84145882	1	0.05	0	0
A	LB	a	19.8069	3.84145882	1	0.05	8.57E-06	0.00030842
A	LC	c	36978.1452	3.84145882	1	0.05	0	0
A	LC	f	597.7171	3.84145882	1	0.05	5.25E-132	1.8908E-130
A	LC	a	3718.5405	3.84145882	1	0.05	0	0
V	CB	c	33399.6299	3.84145882	1	0.05	0	0
V	CB	f	1643.8995	3.84145882	1	0.05	0	0
V	CB	h	206.3645	3.84145882	1	0.05	8.53E-47	3.07162E-45
V	CS	c	33793.2204	3.84145882	1	0.05	0	0
V	CS	f	1613.6851	3.84145882	1	0.05	0	0
V	CS	h	238.0679	3.84145882	1	0.05	1.04E-53	3.73507E-52
V	LB	c	32928.8764	3.84145882	1	0.05	0	0
V	LB	f	1530.8445	3.84145882	1	0.05	0	0
V	LB	h	29.3328	3.84145882	1	0.05	6.10E-08	2.19437E-06
V	LC	c	36155.3416	3.84145882	1	0.05	0	0
V	LC	f	766.7778	3.84145882	1	0.05	9.02E-169	3.2484E-167
V	LC	h	2690.2346	3.84145882	1	0.05	0	0
E	CB	c	31134.9186	3.84145882	1	0.05	0	0
E	CB	f	1836.7874	3.84145882	1	0.05	0	0
E	CB	w/h	588.5762	5.99146455	2	0.05	1.56E-128	5.6054E-127
E	CS	c	31709.1751	3.84145882	1	0.05	0	0
E	CS	f	1821.0214	3.84145882	1	0.05	0	0
E	CS	w/h	671.9159	5.99146455	2	0.05	1.25E-146	4.4835E-145
E	LB	c	31021.6716	3.84145882	1	0.05	0	0
E	LB	f	1782.5846	3.84145882	1	0.05	0	0
E	LB	w/h	610.5064	5.99146455	2	0.05	2.69E-133	9.6945E-132
E	LC	c	39040.6578	3.84145882	1	0.05	0	0
E	LC	f	1170.3051	3.84145882	1	0.05	1.73E-256	6.2401E-255
E	LC	w/h	6936.2359	5.99146455	2	0.05	0	0