Mental representation of equivalence and order

Ulrich von Hecker[1], Elisabeth Müller[2], Stefan Kirian Dill[2], and Karl Christoph Klauer[3]

[1]School of Psychology, Cardiff University, UK.

[2]Universität Mannheim, Germany.

[3]Albert-Ludwigs-Universität, Freiburg, Germany.

Author Note

Correspondence should be addressed to Ulrich von Hecker, School of Psychology, Cardiff University, Tower Building, Park Place, Cardiff, CF10 3AT, United Kingdom. Electronic mail may be sent to vonheckeru@cardiff.ac.uk.

https://osf.io/28fhp/?view_only=51c80d5658754e09b10f879654060a6d

Abstract

With mental models based on relational information, the present research shows that the semantics expressed by the relation can determine structural properties of the constructed model. In particular, we demonstrate a reversal of the classical, well-replicated Symbolic Distance Effect (SDE), as a function of relational semantics. The classical SDE shows that responses are more accurate, and faster, the wider the distance between queried elements on a mentally constructed rank order. We replicate this effect in a study using a relation that expresses a rank hierarchy ("older than", Experiment 4). In contrast, we obtain a clear reversal of the same effect for accuracy data when the relation expresses a number of equivalence classes ("is from the same city", Experiments 1 - 3). In Experiment 3 we find clear evidence of a reversed SDE for accuracy and latency in the above standard condition, and flat curves of means, across pair distances, for accuracy and latency in a condition that makes equivalence classes salient from the beginning. We discuss these findings in the context of a process model of equivalence class formation based on learned piecemeal information.

182 words

Key Words: Analog representations, spatial processing, linear orders, equivalence relations

Word Count: 8527

https://osf.io/28fhp/?view_only=51c80d5658754e09b10f879654060a6d

Introduction

The Symbolic Distance Effect (SDE) is a classic, and one of the most robust and most often replicated effects in experimental psychology. Generalizing across a range of paradigm variants, the effect shows that after learning a transitive order relation between entities (objects, fictitious persons, etc.), for example, A is older than B, B is older than C, C is older than D, … etc., participants are more accurate (and quicker when accurate) reacting to tested pairs (e.g., who is older?) of wider distances on the order sequence A, B, C, D, … (e.g., AD) compared to shorter distances (e.g., AB, De Soto et al., 1965; Foos & Sabol, 1981; Kalra et al. (2020); Kumaran & McClelland, 2012; Leth-Steensen & Marley, 2000; Pohl & Schumacher, 1991; Potts, 1972, 1974; Smith & Foos, 1975; Smith & Mynatt, 1977; Trabasso & Riley, 1975; Trabasso, Riley & Wilson, 1975, Wu & Levy, 2001). In much of this research the distance effect is seen as a marker of analog magnitude processing.

Formatted: Font: Not Italic

Formatted: Font: Not Italic

In this article, we report a reversal of this robust effect within the same general paradigm when using equivalence relations, for example, *A is in the same class as B, B is in*

the same class as C, etc., that is, decreasing accuracies and increasing response times (RTs) for correct responses, as a function of increasing pair distance. We also suggest that the direction of the SDE ("normal" or "reversed") is determined by the relational semantics involved in the task, implicating either a hierarchical order (normal SDE) or an equivalence class (reversed SDE). We first demonstrate that the SDE reverses when an equivalence relation (as opposed to an order relation, Experiments 1 and 2) is to be learned. We then predict and find (Experiment 3) that in the case of a reversed SDE, the reversal itself can be mitigated or even nullified, leading to flat curves of accuracy and latency means as a function of pair distance, by having participants mentally construct category labels for the stimuli, enhancing the learning of equivalence classes. Finally, we also provide a replication of the "normal" SDE, using the same basic paradigm, when an order relation is to be learned (Experiment 4). We discuss these findings in the context of assumptions about how spatial cognition might contribute to the representation of relations of different types.

There are a number of reasons for asking questions about the conditions determining the robustness of the SDE, and the theoretical assumptions underlying its emergence.

1. *Theoretical*: The SDE has often been interpreted as suggesting a spatial representation of the learned order. The idea here is that the learned, individual pieces of pairwise information, for example, *A is older than B*, *B is older than C*, etc., are integrated into a mental representation as a spatial dimension representing the quantity in question, in this case, *age*. Following this assumption, it has been argued that wider distances on this dimensional representation could be better discriminated than narrower distances (Holyoak & Patterson, 1981; Huttenlocher, 1968), therefore engendering higher levels of accuracy, and speedier correct responding, to wider than narrower test queries. Although the basic SDE can be modelled on the basis of non-spatial assumptions, empirical evidence for the contribution of spatial processes has been repeatedly demonstrated (Hinton et al., 2010; von Hecker et al.,

2016, 2019; see also von Hecker & Klauer, 2021). If one assumes that a quantifiable, transitive order (e.g., *age, wisdom, wealth,* etc.) is projected onto a representational space (an extended linear dimension), then this begs the question to what extent the emergence of an SDE (as a signature of that projection) is tied to just this type of representational matching. In other words, given the semantics of the learned relation changing for example from quantity to identity, whilst still transitive, the conjecture is that an SDE would not necessarily emerge. Consider that strict order relations are asymmetric: a R b implies not(b R a). In contrast, equivalence relations are symmetric: a R b implies b R a. Asymmetry supports a representation as linear spatial sequence in case of order relations, in which case the relative left-right positioning between any two elements represents information about the relation between them. In contrast, linear spatial sequences are not privileged vis-a-vis other possible spatial arrangements (e.g., as two-dimensional clouds or cliques of points) in case of equivalence relations. It is possible that therefore, one would not necessarily even attempt to form a linear spatial representation in this case. If so, when later tested on any two elements, pairwise reconstruction based on the learned piecemeal information would be a default strategy. A reversed SDE is then predicted because when a distance of more than one step is queried, the response also requires more than one piece of pairwise learned information to be aggregated (e.g., AB and BC if the distance AC is being queried). This implies a higher overall probability of at least one error being made the more steps that have to be covered, during retrieval of the individual piecemeal information. This again implies decreasing accuracy with the number of necessary steps.

2. *Empirical / Same paradigm, individual differences*. In a previous study (Sedek & von Hecker, 2004) we demonstrated a reversed SDE (for accuracies, not latencies) using the classical SDE-producing paradigm described above (only stepwise pairs AB, BC, CD were presented for learning). The study compared a non-depressed sample and a sample in states

of subclinical depression. For the latter population, as we had argued, integration of piecemeal information into a more complex mental representation was more difficult (Kofta & Sedek, 1998; for a review: von Hecker, Sedek, & Brezicka, 2013). We found an SDE replication for accuracies in the non-depressed control group, but a reversed SDE in the depressed group, such that participants' accuracy *declined* with pair distance. This was seen as showing that the non-depressed group might have used transitive reasoning during learning in order to construct an integrated mental representation, whereas the depressed group, finding it difficult doing the second step (integration), would have to use transitive reasoning only at the time the test items were presented rather than constructing a mental representation beforehand. In particular, we assumed that for a given test relation, participants might have had to remember all stepwise pairs needed to construct an answer deductively "on the spot" (Gilhooly, 1998) rather than reading the answer off a pre-compiled, integrated model. Assuming this process leads one to predict a reversed SDE in accuracies as errors might occur at any given point in the deductive (transitive) chain. At a more general level, these results mean that the SDE may be more malleable than its overall robustness suggests, so therefore the question arises whether transitive reasoning on the basis of the learned piecemeal information might be a task strategy not only characteristic of a particular mood disposition, but possibly also potentially triggered by task constraints. Equivalence relations, as they do not imply linear quantification between elements, might have to be constructed in a way different from projection onto a linear spatial dimension. Learning of this type of relation might therefore engender a reversal of the SDE, as transitive reasoning about equivalence relations may not necessarily use the same type of spatial projection.

3. *Empirical / Different paradigm*. The literature on stimulus equivalence class learning (SE), although not explicitly connecting to the literature on SDE, reports its mirror image with some consistency: decreasing accuracies (and increasing latencies for correct answers)

as a function of pair distance, (Bentall, Jones, & Dickins, 1999; Fields & Verhave, 1987; Fields et al., 1990, 2012; Imam, 2001; Kennedy, 1991). The emphasis of interest is different between our approach and SE. In SE, the emphasis is on investigating the formation of equivalence classes over sets of relations amongst content- and context-deprived stimuli. Therefore, because the stimuli "do not […] share any physical properties, their relatedness is very likely to be the result of training, either in the laboratory or in a natural setting." (Fields & Verhave, 1987, p. 317). In contrast, our emphasis is on the mapping of a relation with a pre-existing, colloquial meaning onto a spatial dimension. It is precisely our interest in the *a-priori* existing meaning that makes us think that relational content may have a determining influence on the type of representation being formed, and the reasoning processes that are engaged when learning such relations. This difference in emphasis is consequential in terms of used paradigms: SE learning does not use the paradigm described above, but mostly proceeds following some variant of the "matching-to-sample" procedure: Stimuli A, B, C, and D, e.g., four abstract drawings with no planned *a priori* connection between them, are learned by each of them (e.g., A) being presented as "sample" vis-à-vis a table displaying some comparison stimuli, one of them being the equivalence target (e.g., B). A participant correctly selecting B from the comparison table is positively reinforced. In this way, chains of stimuli can be trained that would form an equivalence relation. In later tests, stimulus pairs of different pair distance (see above) may be queried, in some cases (see the studies mentioned above) revealing decreasing accuracy and increasing response latencies for correct responses with increasing pair distance; that is, a reversed SDE. This pattern has been explained under the notion of "nodal distance effect" as showing the effect of intervening numbers of stimuli between the two elements of a queried pair (Spencer & Chase, 1996; Fields et al., 1993; but see Imam, 2001, for the argument that the emergence of this nodal effect might depend on chosen methodology). Note that equivalence learning in this way only requires an

understanding of the formal semantics of the relation (i.e., "is equivalent to"), as opposed to surplus semantics in terms of what meaning the relation conveys outside its formal structure, (e.g., "*age*" when learning the relation "*older than*"). In the present approach we integrate the approaches from both literatures, arguing that it is precisely the surplus relation semantics that triggers the processes involved in forming a mental representation of the stimuli, and therefore determine whether a standard SDE or a reversed SDE will be obtained.

Experiment 1

In a first study, we accommodated the classical, SDE-inducing, order learning paradigm (e.g., Leth-Steensen & Marley, 2000; Potts, 1974) for use with an equivalence relation. This meant basically to use a colloquial semantics as a rationale for the relation, in this case: fictitious persons coming from three different cities. This allowed us to present pairs of these persons by the relational connective "is from the same city as". From the above considerations, and in parallel to the findings in the SE learning literature, we predicted decreasing accuracy and increasing response latencies as a function of pair distance (*reversed SDE*).

Method

Participants

Data collection took place simultaneously, online at Cardiff University (Experiment 1a), as well as face-to-face in a lab at the University of Freiburg (Experiment 1b). At the respective places, materials and instructions were presented in English or German.

Experiment 1a

40 students from Cardiff University took part in the online experiment. They received course credit as compensation for their participation. They had mostly English-spoken backgrounds (31 people reported English as their first language), two people named Welsh,

another two Chinese, one Bulgarian and one Spanish as their first language. The majority of participants (27 people) was female, 9 people identified as male and one as non-binary. The mean age was 19.35, with a range from 18 to 23. No participant was excluded.

Experiment 1b

In Freiburg, 30 people were recruited for a laboratory-based study. They received course credit or 5€ for their participation. All except one participant reported German as their native language. 70% of this sample were women, 9 people identified as male. The mean age was 23.9, with a range from 19 to 44. Additionally, participants' occupation or field of study were collected, with only 5 people naming psychology as their field of study. No participant was excluded. This means that the sample collected in Freiburg is more heterogenous regarding age as well as occupational background than the Cardiff sample where only young students of psychology were recruited.

Materials (both 1a and 1b)

In each block, a set of nine names was randomly chosen out of a large pool of names, matched for frequency of recent use (German names for data collection in Freiburg, see von Hecker, Klauer and Aßfalg, 2019; British names for data collection in Cardiff, see von Hecker, Klauer, Wolf and Fazilat-Pour, 2016). Three blocks used only female names, the other three only male names. Out of these nine names, three chains of names were constructed: the main chain with five elements (hereafter denoted as ABCDE), and two secondary chains with two elements each (hereafter denoted as FG and HI). In the learning phase, name pairs consisted of two neighboring names in each of the chains, so that only pairs with a distance of 1 step were presented (all six pairs in the learning phase: AB, BC, CD, DE, FG and HI). The order of names within each pair was however randomized.

For the test phase, name pairs were created by determining all permutations of two names within the chains (e.g. AB, AC, AD, AE, … , ED, EC, EB, EA and FG, GF → 24

pairs), as well as the combinations of all names from the secondary chains with all names of the main chain (e.g. FA, FB, FC, …, GD, GE → 20 pairs), and the combinations of all names of the secondary chains with each other (e.g. FH, FI, GH, GI → 4 pairs). This resulted in a total of 48 name pairs for the test phase, 24 of these being two names from the same city and 24 being two names from different cities.

Names were presented in white font on a dark gray background. The font size was set to 8% of the screen height.

Procedure (both 1a and 1b)

First, participants were asked to fill in a demographics form. In the Freiburg version of the experiment, this consisted of age, gender, occupation/subject of study, handedness, native language and the current presence of vision problems. In the Cardiff version, participants were only asked for their age, gender and native language.

Participants were then instructed that they would be shown pairs of names from the same city. Their task was to memorize which people were from the same city, so that they could later decide if two names were from the same or from different cities. The experiment had six blocks that each consisted of a learning phase and a testing phase. In the learning phase, participants were shown six different pairs of names with four repetitions each, which resulted in a total of 24 learning trials. The sequence of pairs was determined randomly for each of the four cycles, but each pair had to occur once before a pair could be repeated. In each learning trial, two names were shown one above the other, separated by the text "is from the same city as" in the middle of the screen. This was presented for 4 s, followed by a blank screen inter-stimulus interval of 2 s. The order of names within each pair was also determined randomly for each trial.

The testing phase followed immediately after each learning phase. Participants were instructed to decide if the two shown names were *from the same city* or *from different cities*.

They were then presented with 48 test trials in random sequence. Each test trial began with a 1 s fixation stimulus ("x"), after which a pair of names was presented along with a reminder of the key mapping in the right corner of the screen. Now, the names were aligned horizontally next to each other with a gap in between. The left-right position of names within the pair was determined randomly. Participants indicated their judgment by pressing either the left or right arrow key. The key mapping was counterbalanced between participants. After the response was given in an open response interval, a 2 s blank screen inter-stimulus interval followed before the next trial began. The experiment lasted about 25 minutes.

Results

Comparability of the two experiments 1a and 1b

As Experiment 1a being an online study and Experiment 1b a laboratory-based study, as well as 1a being conducted in Cardiff and 1b in Freiburg, a first concern was about comparability in terms of the main dependent variables, accuracy and response latency. We ran preliminary analyses to find that overall accuracy was significantly higher in Freiburg (84%) than in Cardiff (70%), $p < .001$, and that average response latency for correct responses tended to be shorter in Cardiff (1.50*s*) than in Freiburg (1.79*s*), $p = .09$. This pattern, resembling a speed-accuracy trade-off, is unsurprising given the practical constraints prevailing in online versus laboratory-based experiments. Importantly however, there was no interaction between the factors pair distance and place, either on accuracy ($p = .85$) or latency ($p = .39$), such that both datasets were combined (yielding a total N=70) for the more detailed analyses reported below.

General approach to data analysis

For a detailed descriptive presentation and visualisation of accuracy and latency data across all reported experiments see the Supplement file on OSF. The accuracy and latency data sets of all reported experiments in this article were each analyzed in two steps. We

estimated linear mixed models (for the accuracy data: generalized linear mixed models with logistic link function) with participants as random factors, and first determined which random structure would best fit the data. Subsequently, a final model with appropriate random effects was used to evaluate fixed effects (see Jaeger, 2008; Judd, Westfall, & Kenny, 2012). The strategy for selecting a model with appropriate random-effects structure is described in Appendix A, along with information about the particular random-effect structure adopted for each model in each experiment. Effect sizes (Cohen's d_z) are reported across all experiments for those effects that are interpreted as relevant to the main hypothesis, that is, the effect of the factor pair distance. The independent variables were effect-coded as factors.

Accuracy

The average accuracy was 76% in the combined dataset. There were three types of trials in the design: a) trials presenting two persons from different cities, b) trials presenting two persons from the same city with pair distance = 1 which is identical to the learning material, and c) trials presenting two persons from the same city with pair distances greater than 1. Only the latter trial type (c) is being used to evaluate the hypothesis. Trials of type a) are not linked to the hypothesis, and accuracies on trials of type b) would represent a confound between memory performance on the basis of just-presented material and possible retrieval from an integrated model. Accuracy data of all above types of trials can be seen in Table 1.

Excluding trials of type a), the final model to be evaluated had fixed effects for pair distance (1 step, 2 steps, 3 steps, 4 steps), block number (1...6) and the interaction. We found significant effects for pair distance, $\chi 2(3) = 11.67$; $p = .009$ and block number, $\chi 2(5) = 14.89$; $p = .011$. The interaction was not significant. Inspection of the accuracy means revealed the highest level at pair distance 1 (.81) and a stepwise decline (.71, .69) to the level attained at pair distance 4 (.67), see Table 1, thus confirming a reversed SDE. With respect to the main

hypothesis, we found a significant linear contrast covering the pair distances pertaining to not-presented pairs, that is, distances hypothetically existing on the linear mental model (2, 3, and 4), and as such omitting pair distance 1 (type b) trials, see above, $z = 2.74$, $p = .006$, $dz = .25$.

For the block number factor, a planned contrast revealed a linear increase in accuracy from block 1 to block 6, $z = -3.766$, $p < .001$, most likely due to a practice effect across the duration of the experiment.

Latency

For correct responses, latency data from this and the following experiments were trimmed according to the Tukey criterion based on excluding outliers with values larger (smaller) than the upper (lower) quartile plus (minus) 1.5 times the interquartile range in the individual's distribution of latencies (see Clark-Carter, 2004, Chapter 9). Response time analyses across all experiments were repeated with the untrimmed, log-transformed data, and yielded the same results as reported below. The final model had the same fixed effect structure as the one above for accuracy. With respect to our hypothesis, the picture is mixed. We did find a significant effect for pair distance, $F(3, 139.70) = 6.16$; $p < .001$, showing an overall tendency of these means to increase across pair distances (see Table 1) which is in line with a reversed SDE. However, addressing the hypothesis more specifically with planned contrasts across inferred pair distances 2, 3, and 4, we did not find evidence for a linear trend, $d_z = .001$, but instead a tendency for a quadratic trend, $z = 1.624$, $p = .10$. However, in terms of latencies, authors have made a point with regard to the widest distance, that is, the pair that contains both end elements. Such a pair is likely to be privileged for relatively quick responses because individuals can respond as soon as they have identified one of the two elements as maximum or minimum (Potts, 1972, 1974). Alternatively, semantic codes may be generated easier for stimuli at the end points, leading to faster responses to the pair

representing the widest distance (Shoben et al., 1989; Leth-Steensen & Marley, 2000). Therefore, we also ran a contrast excluding the widest pair, containing both end elements, and found a tendency of slower responding for pair distance 3 in comparison to pair distance 2, $z = 1.753$, $p = .07$.

For the block number factor, a planned contrast revealed a linear decrease in response latency from block 1 to block 6, $z = -2.58$, $p = .004$, most likely due to a practice effect across the duration of the experiment.

Discussion

In a first experiment, we received support for a reversed SDE in accuracies when using equivalence relations in the classic order learning paradigm, using colloquial semantics. Accuracies in responding to pairs that had not been presented, but had to be inferred by the participant, yielded a stepwise decline from step 2 to step 4 of pair distance. This supports the assumption that participants may have inferred their response via transitive reasoning rather than retrieved a response from an already-compiled linear mental model. Using this strategy, the more elements are transitively included in the calculation in order to generate a response to a pair query, the more likely it is that an error occurs, so accuracy will decline with pair distance. However, in terms of response latencies the picture from this experiment is not conclusive, although the absence of a classic SDE in these data suggests that processes different from those associated with order relations (see above) are underlying the observed pattern. Although we have some hint on an increase with more elements included in the query, the results may be confounded by other factors, for example, pair distance 4 representing both end elements. This pair has been shown as privileged for fast responding in other studies (see above). Experiment 2 was designed to address this confound.

Experiment 2

We aimed at a replication of the SDE reversal as obtained in Experiment 1 under conditions of equivalence class learning. Additionally, we aimed at ruling out one particular explanation for why, in Experiment 1, the reversal was clearly observed only in terms of accuracies but not response latencies. This explanation leans on the fact (see Table 1) that the widest pair distance (4 steps), was associated with a quicker RT than expected when assuming, on the grounds of an SDE reversal, a linear RT increase as a function of step distance. What distinguishes pair distance 4 from all other pair distances is that it contains both end elements. As mentioned above, end elements may be privileged in terms of response speed (Leth-Steensen & Marley, 2000; Potts, 1972, 1974; Shoben et al., 1989), so the reason for the observed RT being shorter than expected could have to do with this particular privilege. Therefore, as a rationale for the present experiment, we changed the design to be able to observe responses vis-à-vis pair distance 4, without that pair being composed of both end elements.

Method

Participants

Data collection was online at Cardiff University with $N = 42$ participants (1 male) with English-spoken backgrounds. Three participants were excluded for RT analyses, by Tukey criterion (see above), yieding $N = 39$. Mean age was 19.0 years, the range was 18 to 28. Participants received course credit for their participation.

Materials and Procedure

The same materials and procedures were used as in Experiment 1, with the following exceptions: 1. Only four blocks were used, two populated with female names, and two with male names.

2. In the learning phase, there were now ten names in each block (instead of nine as in Experiment 1). The main chain now contained *six* instead of five names as elements

(XABCDE), with the first name (X) included in the learning phase but not in the testing phase. The secondary chains (FG and HI) remained unchanged. In the learning phase, name pairs consisted of two consecutive names in each of the chains, but now starting with X, not A (as in Experiment 1), so that only pairs with a distance of 1 step were presented (all seven pairs in the learning phase: XA, AB, BC, CD, DE, FG and HI). As in Experiment 1, each pair was presented four times during the learning phase (28 pairs in total). The order of names within each pair presentation was randomized.

The testing phase was identical to Experiment 1. By this measure, the learned end elements in the main chain were X and E, but the widest pair to be tested was AE as before, but now not representing both end elements[1]. The experiment lasted approximately 25 minutes.

Results

Accuracy

The average accuracy of the combined dataset was 67% which is lower than in Experiment 1, possibly due to increased difficulty, as one more person was included in the learning phase. Accuracy means are displayed in Table 2. The final model had fixed effects for pair distance (1 step, …, 4 steps), block number (1,…, 4) and the interaction. Block number had a significant effect, $\chi^2(3) = 16.25$; $p < .001$, with a planned contrast revealing block 1 being at a lower level of accuracy than the later three blocks which all were at similar levels, $z = 3.355$, $p < .001$. Pair distance yielded a significant effect, $\chi^2(3) = 17.48$; $p < .001$, with a significant linear trend showing declining accuracy across the four pair distances, $z = 4.474$, $p < .001$. Focusing on the three non-presented pair distances 2, 3 and 4, there was a significant linear trend across these three levels, $z = 2.305$, $p = .02$, $d_z = .33$. These results constitute again a reversal of the SDE, and a replication of Experiment 1.

Latency

A model with identical fixed effect structure as above was statistically evaluated. It yielded no significant effects.

Discussion

In the second experiment using equivalence classes, the SDE reversal was again found for accuracies. Results for response latencies were inconclusive, as such not supporting the assumptions about combined end point effects that had led to the minor paradigm change in this experiment[2]. However a flat curve across pair distances for RT at least does not support the existence of a spatial order representation of chain elements as the latter would imply a classical SDE: Patterns of response time data in linear order learning, in the past, have very often been coupled with corresponding accuracy data patterns to show a decrease in response latency whilst accuracy simultaneously increases, across pair distances (Leth-Steensen & Marley, 2000; von Hecker & Klauer, 2021).

In the SE learning literature, however, there is one specific observation which leads to a heuristic corollary to be followed up from the present experiment: The observed pattern of increased RTs across pair distances appears to be sometimes weakened (the linear contrast being insignificant or getting weaker) in cases where participants had been given names for the to-be-learned equivalence classes up front (Bentall et al., 1993; but see Fields et al., 2012). Or, a similar attenuation of the "nodal effect" on *accuracies* was observed in later phases of testing, that is, when participants had had more extended opportunities to learn and rehearse the materials (Spencer & Chase, 1996). In these later phases of testing, participants responded accurately to all pairs regardless of the nodal distance they represented.

It is possible that such increase in practice would strengthen the mental representation of equivalence classes during the process of them being formed (as entities). The above group of findings could therefore be interpreted as showing that extended practice or semantic facilitation may result in a more clearly emerging class structure. If so, one may conclude that

the link between each learned element (within a transitive chain) and the class concept should eventually be exactly one step only, as class membership would have been consolidated as a feature of that element proper. The result of such a consolidation process would be a flat curve in terms of accuracies and RTs. To investigate this idea, a comparison between a condition replicating the above paradigm, and another condition in which the to-be-formed classes are already highlighted explicitly during learning, should be informative.

Experiment 3

The methodology of Experiment 1 was replicated, now using two conditions: In condition "no colour", both names in the learning phase, shown vertically, were presented in white font. In condition "colour" three colours were used as fonts, one for each of the three to-be-formed equivalence classes. In the learning phase, the vertically presented names appeared in the respective colour of their class. As such, colour could be used as a label for the equivalence class that a given pair was in. We expected a dissociation for accuracies and/or response latencies such that in the "no colour" condition a reversed SDE will occur, but flat curves for both dependent variables in the "colour" condition.

Method

Participants

Data collection was lab-based at Freiburg University with $N = 44$ participants (13 male) with German-spoken backgrounds. Mean age was 24.4 years, the range was 18 to 44. No participant was excluded. Participants were randomly assigned to one of two groups, that is, either condition "no colour" ($N = 23$) or condition "colour" ($N = 21$). They received course credit or 5€ for their participation.

Materials and Procedure

The same materials and procedures were used as in Experiment 1, except in condition "no colour" all names in the learning phase were presented in white font, whereas in condition "colour" each name in the learning phase was presented in a font specific to the class they were in. That is, the colours blue, green, and orange were randomly assigned to signify the main chain and the two secondary chains within the learing material. Each name in one of these chains was presented in its corresponding colour. Keeping instructions identical between "no colour" and "colour" condition, colours were not mentioned to participants in the latter condition. In the test phase all names were presented in white. The experiment lasted about 25 minutes.

Results

Accuracy

The average accuracy was 85%. Accuracy means are displayed in Table 3. The final model had fixed effects for color version (no colour vs. colour), pair distance (1 step, …, 4 steps), block number (1,…, 6), and the respective interactions. Block number had a significant effect, $\chi^2(5) = 16.51$; $p = .006$, a linear contrast revealing a significant improvement of accuracy across the six blocks, $z = 4.15$, $p < .001$. The only further significant effect was the interaction between color version and pair distance, $\chi^2(3) = 10.74$; $p = .013$. Further investigating this interaction, we ran separate follow-up analyses for the two groups (no colour vs. colour). In the "no colour" group, there was a significant effect of pair distance, $\chi^2(3) = 12.50$; $p = .006$, showing a reversed SDE in that accuracies decreased from pair distance 1 to pair distance 4 (see Table 3). In particular, the linear contrast only comprising the three inferred pair distances (2, 3, and 4) was also showing a significant decrease in accuracy, $z = 3.596$, $p < .001$, $d_z = .59$.

On the other hand, in the "colour" group, the separate analysis yielded no significant effect, as the level of accuracy appeared equal across pair distances (see Table 3), especially when considering a contrast only comprising the three not-presented distances (2, 3, and 4), z

= -0.917, p = .359. This linear contrast was also significantly smaller than the corresponding one in the colour group, z = 2.394, p =.017. As explained above, responses of types a) pertaining to two names from different cities were not considered for analysis, and accuracies of responses of type b), pair distance 1, may be inflated by the fact that these very relations had just been presented in the learning phase.

Latency

A model with the same fixed effect structure as for accuracies was run for response latencies. There were overall significant effects of pair distance, $F(3, 55.92) = 3.43$; p = .02, and block number, $F(5, 66.89) = 6.22$; $p < .001$, the latter showing a linear decrease in response latencies across blocks, presumably due to practice (see Table 3). No other effects were significant; in particular, the interaction between colour version and pair distance yielded a tendency, $F(3, 55.92) = 2.25$; p = .09. Following up on this tendency, we ran separate models for both condition groups. In the "no colour" group, following a significant pair distance effect, $F(2, 1294.16) = 9.98$; $p < .001$, latencies increased across pair distances 2, 3, and 4, as a linear contrast revealed, t = -2.40, p = .02, demonstrating a reversed SDE across the distances covering inferred relations, d_z = .46.

In the "colour" group, the pair distance effect remained insignificant, $F(2, 986.46)$ = .99, p = .372, confirming the flat (or even seemingly decreasing) curve of latencies as seen in Table 3, across pair distances 2, 3, and 4.

Discussion

The results of the manipulation used in the present experiment can be interpreted as follows. The process of learning and integrating a number of pairwise relations into a representation of equivalence classes eventually produces a representational model consisting of separate classes that contain a number of elements each (see also Bentall et al., 1999; Fields et al., 1993). In each class, the elements are not articulated (e.g., ordered) amongst

each other in any way, therefore the link to the class concept may be supposed to be equally strong (step length = 1) for all elements, without hierarchy. In order to arrive at such a model in the first place, the incoming initial pieces of information (individual relations about "who is from the same city") have to be processed in terms of transitive reasoning, producing a reversed SDE. This means, as the learning of the material proceeds from taking in one piecemeal relation to taking in the next one, a chain is constructed that has end points and varied distances between its elements, although the magnitude of such distances is meaningless. Still, the chain implies longer latencies for queries about pairs of elements that lie more distanced on the chain than others, e.g., AE in comparison to AB. This also implies more errors being made along the chain, the more distant two queried elements lie on the time line, as probabilities of committing errors rise with respect to retrieval or connectivity, the more elements involved in the reasoning chain[3].

In the "no colour" condition, this stage of the constructive process is mirrored in terms of a decrease in accuracy and an increase in latency as a function of increasing pair distance. In contrast, in the "colour" condition, we assume that this manipulation facilitates and speeds up the constructive process into a stage where class labels attached to individual elements are already well accessible. In this situation, we find equal levels of accuracy and no difference in response latencies across pair levels, reflecting an already stronger formed mental class model, as there would be an equal distance of each element to the class concept[4].

The latter interpretation needs to be treated with caution insofar it rests on a null result in the "colour" condition. Thus, we cannot rule out the existence of a small distance effect, of a size that we were not able to capture with the level of test power realised in this experiment. Nevertheless, the distance effect in the "colour" condition was significantly smaller than the distance effect in the "no colour" condition in support of our interpretation.

Experiment 4

The semantics of the relation is predicted to determine the kind of mental model construction that takes place. Experiments 1 to 3 yielded a consistent demonstration of a reversed SDE in accuracies when an equivalence relation was used. The present experiment serves as a comparison in terms of using an order relation. In this case, we predict that within the same basic paradigm as before, a linear representation of the order dimension is formed, arranging the elements into a dominance hierarchy on the semantic dimension, thereby producing the classical SDE (Leth-Steensen & Marley, 2000; Potts, 1972, 1974; Smith & Foos, 1975; Smith & Mynatt, 1977; Trabasso & Riley, 1975).

Method

Participants

Data collection was simultaneously carried out, laboratory-based at Cardiff University ($N_{\text{Cardiff}} = 22$), all with English-spoken background, and laboratory-based at Freiburg University ($N_{\text{Freiburg}} = 22$), all with German-spoken background (all materials and instructions were presented in the respective background language at both places). Participants received course credit (Cardiff and Freiburg) or 5€ (Freiburg) for their participation. One participant at Cardiff was excluded due to equipment problems during the session, two participants at Freiburg were excluded, one due to faulty data registration, and one due to the participant terminating the session prematurely. With no other exclusions, the total sample had $N = 41$ participants, 6 of whom were male and 35 were female. The mean age was 22.2 years with a range of 18 to 40 years.

Materials and Procedure

We tried to keep materials and procedures as identical as possible to Experiment 1, but because of the different nature between the two types of relations, the following changes were unavoidable. During the learning phase, name pairs still consisted of two neighbouring names in each of the chains. However, because of the asymmetry implied by order relations,

in a random half of the six pairs in each of the four repetitions, the upper name was the dominant one and the two names were divided by the comparator "is older than", and in the other half, the lower name was dominant and the comparator used was "is less old than". The order of trials was randomized for each repetition (24 learning trials in total, as in Experiment 1). In the test phase, participants were instructed to indicate their judgment on which of the two names represented the older person, by pressing either the left or right arrow key, corresponding to the name they recognised as, or they had inferred to be, "older". The name pairs in the test phase consisted of all combinations of two names within the chains, each pair presented twice, that is, once with the dominant name on the left, and once with the name on the right side in the pair, which resulted in 24 pairs. This set was presented twice to keep the number of test trials equal to Experiment 1 (= 48 test trials). The order of trials was randomized for each repetition. Unlike in Experiment 1, there were no combinations of names between the chains, because it was not possible to infer any information about the order relations between the chains. The experiment lasted about 25 minutes.

Results and Discussion

Comparability of the two samples in Cardiff and Freiburg

As in Experiment 1, we ran preliminary analyses to find that overall accuracy was significantly higher in Freiburg (83%) than in Cardiff (72%), $p < .001$, but there was no interaction between place and pair distance, $p = .36$. For latency, we found no overall main effect of place, $p = .79$, but there was an interaction between the factors place and pair distance, $F(3, 70.89) = 3.69$; $p = .006$. This interaction was due to the fact that latencies for pair distance 1 (relations presented during learning) were quicker in Cardiff (1.60*s*) than Freiburg (1.70*s*), but for pair distances 2, 3, and 4 (inferred relations) quicker in Freiburg (1.71*s*, 1.56*s*, and 1.46*s*) than Cardiff (1.73*s*, 1.68*s*, and 1.61*s*). Given that these three pair

distance levels carry most of the evidential burden, and that the linear trend for inferred relations appeared the same at both places, we still decided to pool both datasets ($N = 41$) for the more detailed analyses reported below.

Accuracy

The average accuracy was 77% in the combined sample. Accuracy means are displayed in Table 4. The final model had fixed effects for pair distance (1 step, …, 4 steps), block number (1,…, 6) and the interaction between both. Block number had a significant effect, $\chi^2(5) = 28.97$; $p < .001$, a linear contrast revealing a significant improvement of accuracy across the six blocks, $p < .001$. Pair distance had a significant effect, $\chi^2(3) = 24.48$; $p < .001$, showing a strong linear component amongst the three inferred levels of pair distance (2, 3, 4), $z = 4.436$, $p < .001$, $d_z = .65$, see Table 4. This replicates the classic SDE, demonstrating an increase in accuracy with (inferred) pair distance. The level of pair distance 1, as discussed above, is confounded with these pairs having just been displayed as learning items, therefore receiving a retrieval advantage during testing.

Latency

A model with similar fixed-effect structure as above, for accuracy, was run. We found block number to have a significant effect, $F(5, 48.07) = 5.52$; $p < .001$, demonstrating a linear trend in decreasing response latencies across the six blocks, presumably due to practice. Crucially, we also found a significant effect for pair distance, $F(3, 69.17) = 11.65$; $p < .001$, A planned contrast confirmed a strong linear trend for response latencies to decrease across pair distances 2, 3, and 4, $z = 5.463$, $p < .001$, $d_z = .94$, see Table 4, as such replicating the classic SDE.

Both accuracy and latency data from this experiment demonstrate the reverse pattern compared with the pattern observed in Experiments 1-3: The wider the distance on the

hypothetical linear model, the more accurate and fast are the observed responses. The SDE is most likely the result of a linear order representation in mental space, whereby the semantics of the comparator ("older than", "less old than") triggers the construction of a linear hierarchy in mental space that supports the rank order between the learned elements (Holyoak & Patterson, 1981; Huttenlocher, 1968; Leth-Steensen & Marley, 2000; von Hecker & Klauer, 2021). This stands in opposition to the semantics of equivalence relations as used in Experiments 1-3 where we observed a *reversal* of the SDE.

General Discussion

The present research investigated the representational quality of equivalence relations as compared to order relations, via the classic Symbolic Distance Effect (SDE) or its reversal, using the same basic paradigm. The paradigm used is the classic linear order learning technique which relies on semantic content of the ordering relation to be present (e.g., "is older than", see De Soto et al., 1965; Foos & Sabol, 1981; Leth-Steensen & Marley, 2000), as opposed to the usual learning technique in the stimulus equivalence field (SE) which implies the use of only abstract symbolic signifiers in match-to-sample trials, thereby avoiding all possible pre-existing semantics associated with the relation (Bentall, Jones, & Dickins, 1999; Fields & Verhave, 1987; Fields et al., 1990, 2012). Our main argument is that the use of a particular semantics contained in the relation (as implying either a hierarchical order or indeed a number of equivalence classes) can determine whether the classic SDE does result or not. In turn, we argue that the presence of an SDE, or alternatively, a reversed SDE, may be seen as signature of the formal structure that the eventually constructed mental model of the relation will have. To be clear vis-a-vis the above-cited literature on SE, the equivalence relation used here is not purely formal or free of semantic content (as it is mostly within SE research), because the content does imply living in the same city or not. Crucially however,

the equivalence relation used here is just not of the same type that can be mapped onto a linear dimension.

To summarise our results, the classic SDE was found reversed under conditions of learning an equivalence relation using the classic technique employed for linear order learning (see above, Experiments 1, 2, 3). Across studies, this reversal was reliably observed for accuracy, but in one study also for response latency (Experiment 3, "no colour" condition)[5]. The "two-end-elements" argument could be ruled out as explanation for the weaker tendency of the reversal to show up in RTs (Experiment 2). Finally, we replicated the classic SDE in an order semantics condition (Experiment 4). All these experiments used the same basic paradigm in which during a learning phase, pairs of elements only directly connected as direct neighbours in a chain are presented (e.g., AB, BC, CD, etc.), as the minimal condition under which the eventual chain can be constructed via transitivity (for spatial representations of ordered elements, see van Dijck et al., 2013). Pairs of elements spanning wider distances on that chain (providing redundancy but making some constructive effort obsolete) are not presented during learning. As such, in a 5-element chain as used here, later performance on queries about pair distances 2, 3, and 4 are most informative as to the type of model that is being constructed because at testing, pairs of distance = 1 can benefit from being previously used as learning material.

In terms of consistency between our accuracy and RT data, it must be taken into account that in general, the level of accuracy obtained in the present series of experiments using equivalence relations (overall accuracy across Experiments 1-3 = .76) is lower than the level obtained in earlier research, across a series of experiments using the same basic paradigm, but order relations (von Hecker et al., 2016, Experiments 1-7: overall accuracy = .85). Thus, the present task (equivalence classes) appears to be more difficult than the previous one (order relations). In this light, hypothesis-related variance might be expected to

be more pronounced within accuracy than latency data, as this is a tendency to be the case for more difficult tasks (see MacLeod & Nelson, 1984; Wickelgren, 1977). Indeed, the literature on SDE acknowledges that the effect does not always show up in exact parallel between the two types of data (Leth-Steensen & Marley, 2000; von Hecker et al., 2013; see also Schubert, 2005). With this background in mind, it is still problematic that we find, in particular, performance at testing the ultimate distance in the 5-element chain (AE = pair distance 4) to be least reliable, thus being the cause for the hypothesis concerning RT not being statistically borne out consistently (Experiments 1 and 2). Correct responses to the AE pair are "too quick" in Experiments 1 and 2, for being in line with the assumption of a reversed SDE (which stipulates a linear increase in RT). Experiment 2 ruled out one explanation for this irregularity: One could argue that AE is constituted of the two end elements of the chain, as end elements have been assumed to benefit from better discriminability compared to elements within the chain (Moyer & Bayer, 1976; Holyoak & Patterson, 1981; Leth-Steensen & Marley, 2000). Whilst this explanation was not supported, one has to keep in mind that the precision of statistical estimation in this paradigm degrades with increasing pair distance levels. The numbers of available instances for testing within a 5-element chain are, as pair distance levels increase, 4, 3, 2, and 1. Thus, the particular performance level observed for pair distance 4 in Experiments 1 and 2 could be co-determined by imprecision of measurement, or indeed another factor we were not able to rule out.

Whilst Experiments 1 and 2 mainly aimed at a demonstration of the SDE reversal under conditions of equivalence relations, theoretically, results from Experiment 3 most immediately capture the way we conceptualise the connection between relational semantics, emergence of SDE (or, reversal of it), and the structure of the mental representation that emerges from learning. In our experiments, equivalence information during the learning phase was delivered in pairs of names neighbouring within the hypothetical chain. The order

of pairs during learning was randomised. To arrange all pairs into an overall consistent representation means to mentally re-arrange all pairs via transitive reasoning, so that connecting elements between the pairs would yield a transitive continuation of the chain (e.g., AB – BC – CD, etc.). This implies that as long as the classes are not firmly established (i.e., given a name or some meaning), the chain of piecemeal information still represents a transitive linear order. In case of the relation representing *order* semantics, the result of this re-arrangement appears easy to directly map onto a more concrete type of meaningful, transitive, anti-symmetric, dimensional concept (see Lipschutz & Lipson, 1997) representing, for example, age, size, wealth, speed, etc. In particular, according to its spatial interpretation (Dehaene et al., 1993, Holyoak & Patterson, 1981; Huttenlocher, 1968; Trabasso & Riley, 1975, Van Opstal et al, 2008, but see Kenny, 1971) such an order representation normally produces the emergence of an SDE or its variant, a numerical distance effect (Sekuler & Mierkiewicz, 1977; Dehaene et al., 1990, and see the present Experiment 4). In contrast, in case of the relation representing *equivalence* semantics, after transitive re-arrangement there is no easy way to map the logical chain onto a *meaningful* dimension that would likewise arrange the elements, linearly, onto its extension. Rather, class labels have to be generated in addition, implying at this stage transitivity and symmetry, as well as a direct association of each element with the pertaining class label. In the equivalence learning literature, it is well documented that participants' class formation runs through different stages, until all logical properties required for successful class generation are satisfied (Bentall et al., 1993, 1999; Imam, 2001; Sadeghi, 2018). Amongst factors facilitating this process, giving the classes names beforehand, or using meaningful relations, have been mentioned (Bentall et al., 1998; Fields et al., 2012).

In terms of the present Experiment 3, we submit that what we observe in the "no colour" condition is a stage of integration *prior to* the classes as such (or their labels) being

sufficiently salient. Therefore, we observe an SDE reversal, inasmuch as retrieval during the test phase still has to proceed via transitive reasoning. In contrast, what we observe in the "colour" condition is a stage of integration in which classes have already been made salient via the different colours attached to individual names, obviating the need for transitive reasoning at test. In this sense, participants' generation of equivalence classes might have been "boosted" because names could be directly associated with one of the colour labels in memory. As a result, mean accuracy levels across all pair distances appeared flat as compared to condition "no colour", that is, levelling the reversed SDE. We can say the same for RTs, just in case of Experiment 3, as the "no colour" condition indeed yielded a reversed SDE for RTs, whilst condition "colour" revealed no RT differences across pair distances. As a corollary, we suppose that our results of reversed SDE's in Experiment 1 and 2 can be seen in parallel to the mechanism outlined for the "no colour" condition in Experiment 3.

We see the contribution of this research also in the context of the research on paired-associative inference. In this field, alternative accounts of emerging *generalisations* in episodic memory in order to derive new knowledge (which terminology corresponds to our use of the term *mental model*, see Banino et al., 2016) are debated, with the focus predominantly on hippocampal functioning. Within an *encoding*-based perspective, it is argued that on the basis of overlapping information (e.g., A>B, B>C), piecemeal information may be integrated already during learning, into blended representations that facilitate later generalisations (e.g., A>C). Alternatively, within a *retrieval*-based perspective, the initial piecemeal information is deemed to be held in memory separately, and retained until a retrieval situation arises, at which point it is re-activated and then combined into an integrated representation. Whereas in this literature, findings of a negative SDE are taken as indicating a predominance of the retrieval-based explanation (Banino et al., 2016), others argue in favour of the existence of both pathways to generalisation in hippocampal functioning (e.g.,

Zeithamova & Bowman, 2020). In the context of this debate, our contribution suggests that both pathways might exist, and that indeed the world-knowledge-based *content meaning* of the relation in question triggers which one will be likely to function in a particular case. Speculatively, we suggest further that it might be the formal correspondence between relational attributes and attributes within the empirical domain, which will immediately trigger one of these choices, namely, integration during learning. A case in point is the learning of order relations. If, for example, A > B, B > C are introduced as meaning that A be older than B, and B be older than C, then an integration into one linear dimension, representing *age*, easily maps onto existing world knowledge, as age differences between individuals are part of our everyday experience. Such integration is useful in yielding new quantitative information (e.g., the distance between A and C should be at least as large as, or larger than, either the distance between A and B, or B and C), thus leading to a classical SDE via encoding-based integration.

On the other hand, in the case of information presented about A = B, B = C (as meaning that A, B, and C come from the same city), integration into a linear order does not yield new quantitative information. As such, the information does not easily map onto the formal concept of a linear order. New (nominal-scaled) information emerges only after more than one class or clique is recognised as existing across the overall set of relations, such that distinguishing class labels are then generated. We submit therefore that in such a case, a linear chain is just a temporary device, constructed in response to test queries, by processing each piecemeal pair one by one. These individual pairs are retained and later retrieved, during a first stage of representation, to generate responses to queries about class membership of individual elements (thereby leading to a reversed SDE). In a later stage, however, class labels are likely to be generated and associated with these elements, such that retrieval based processing becomes obsolete, as does any representation of distance between elements within

one class. The overall emerging mental representation would then amount to a model with more than one class or clique.

Limitations

In our assessment, the present experiments provide relatively conclusive evidence for the hypothesis that the classical SDE effect can be reversed as a function of the semantics of the transitive relation that is to be learned. They also make some progress in the direction of suggesting a processing account of this phenomenon based on the effect that the reversed SDE effect is weakened in the colour condition of Experiment 3. However, this evidence remains tentative at this point for two reasons: First, the colour condition in Experiment 3 was associated with somewhat lowered overall accuracies, which as suggested by the analyses reported in Footnote 5 may have contributed to weakening the reversed SDE effect (note, however, that accuracy levels in the colour condition were clearly above the corresponding accuracy levels in Experiments 1 and 2, which makes this alternative somewhat unlikely)). Second, as already discussed (see e.g., Footnote 3), there are alternative theoretical accounts for the reversal of the SDE effect that we cannot rule out at this point. Taken together, we believe to have demonstrated the reversal of the SDE effect, but more work remains to be done to evaluate our tentative account of it.

Conclusion

The chosen methodological approach to equivalence class learning is different from the Stimulus Equivalence Learning (SE) literature in two important aspects. First, unlike SE, we are using the same paradigm in order to elicit either a classic SDE, or alternatively, its reversal. This means, we do not use match-to-sample procedures with abstract symbols. This procedure makes good sense in the context of most of SE paradigms as they intentionally wish to exclude from the experimental procedure all traces of pre-existing knowledge about

the relation semantics. Second, and complementary to the latter, we believe that precisely the *inclusion* of semantic factors yields the key to an understanding of equivalence versus order learning in terms of generating different forms of mental representations. On the basis of the four experiments reported here, we suggest that if the to-be-learned transitive relation semantically implies a hierarchical order as overall representation (e.g., "is older than…"), a linear representation will emerge, the signature of which is the classic SDE. On the other hand, if the to-be-learned transitive relation semantically implies a number of equivalence classes as overall representation (e.g., cities as emerging from "is from the same city as …", etc.), the structure of the formed representation is eventually that of a class (or, clique) model (von Hecker, 1997; von Hecker et al., 2013). Its signature is, in a first constructive stage, a reversed SDE. Eventually, in a later stage of more salient class labels, its signature is likely to be flat curves across the distance levels of the initially presented chain of elements, a prediction that we intend to test in future studies.

References

Banino, A., Koster, R., Hassabis, D., & Kumaran, D. (2016). Retrieval-based model accounts for striking profile of episodic memory and generalization. *Scientific Reports, 6,* 1-15.

Bates, D., Maechler, M., Bolker, B. M., & Walker, S. (2015). "Fitting Linear Mixed-Effects Models using lme4." ArXiv e-print; in press, *Journal of Statistical Software*, http://arxiv.org/abs/1406.5823.

Bentall, R. P., Dickins, D. W. , & Fox, S. R. A. (1993). Naming and equivalence: Response latencies for emergent relations. *Quarterly Journal of Experimental Psychology, 46B*, 187-214.

Bentall, R. P., Jones, R. M., & Dickins, D. W. (1999). Errors and response latencies as a function of nodal distance in 5-member equivalence classes. *The Psychological Record, 49*, 93–116.

Clark-Carter, D. (2004). Quantitative psychological research: A student's handbook. New York, NY: Psychology Press.

De Soto, C. B., London, M., & Handel, S. (1965). Social reasoning and spatial paralogic. *Journal of Personality and Social Psychology*, *2*, 513.

Dehaene, S., Bossini, S., and Giraux, P. (1993). The mental representation of parity and number magnitude. *Journal of Experimental Psychology: General, 122*, 371–396.

Dehaene, S., Dupoux, E., and Mehler, J. (1990). Is numerical comparison digital? Analogical and symbolic effects in two-digit number comparison. Jou*rnal of Experimental Psychology: Human Perception and Performance, 16*, 626–641.

Fields, L., & Verhave, T. (1987). The structure of equivalence classes. *Journal of the experimental analysis of behavior*, *48*, 317-332.

Fields, L., Adams, B. J., & Verhave, T. (1993). The effects of equivalence class structure on test performances. *The Psychological Record, 43*, 697–712.

Fields, L., Adams, B. J., Verhave, T, & Newman, S. (1990). The effects of nodality on the formation of equivalence sets. *Journal of the Experimental Analysis of Behavior, 53*, 345-358.

Fields, L., Arntzen, E., Nartey, R. K., & Eilifsen, C. (2012). Effects of a meaningful, a discriminative, and a meaningless stimulus on equivalence class formation. *Journal of the Experimental Analysis of Behavior*, *97*, 163-181.

Foos, P. W., & Sabol, M. A. (1981). The role of memory in the construction of linear orderings. *Memory & Cognition*, *9*, 371-377.

Gilhooly, K. J. (1998). Working memory, strategies, and reasoning tasks. In R. H. Logie & K. J. Gilhooly (Eds.), *Working memory and thinking* (pp. 7–22). Hove, England: Psychology Press.

Hinton, E. C., Dymond, S, von Hecker, U. & Evans, C. J. (2010). Neural correlates of relational reasoning and the symbolic distance effect: Involvement of parietal cortex. *Neuroscience, 168,* 138-148.

Holyoak, K. J., & Patterson, K. K. (1981). A positional discriminability model of linear-order judgments. *Journal of Experimental Psychology: Human Perception and Performance*, *7*, 1283.

Huttenlocher, J. (1968). Constructing spatial images: a strategy in reasoning. *Psychological Review, 75*, 550–560.

Imam, A. A. (2001). Speed contingencies, number of stimulus presentations, and the nodality effect in equivalence class formation. *Journal of the Experimental Analysis of Behavior*, *76*, 265-288.

Jaeger, T. F. (2008). Categorical data analysis: Away from ANOVAs (transformation or not) and towards logit mixed models. *Journal of Memory and Language, 59,* 434-446.

Judd, C. M., Westfall, J., & Kenny, D. A. (2012). Treating stimuli as a random factor in social psychology: a new and comprehensive solution to a pervasive but largely ignored problem. *Journal of personality and social psychology*, *103*, 54.

Kalra, P. B., Binzak, J. V., Matthews, P. G., & Hubbard, E. M. (2020). Symbolic fractions elicit an analog magnitude representation in school-age children. *Journal of Experimental Child Psychology*, *195*, 104844.

Kennedy, C. L. (1991). Equivalence class formation influenced by the number of nodes separating stimuli. *Behavior Processes, 24*, 219-245.

Kenny, A. (1971). "The homunculus fallacy," in G. Marjorie, Ed.: I*nterpretations of Life and Mind: Essays Around the Problem of Reduction*, New York: Humanities Press.

Kofta, M., & Sedek, G. (1998). Uncontrollability as a source of cognitive exhaustion: Implications for helplessness and depression. In M. Kofta, G. Weary, & G. Sedek (Eds.), *Personal control in action: Cognitive and motivational mechanisms* (pp. 391–418). New York: Plenum Press.

Kumaran, D., & McClelland, J. L. (2012). Generalization through the recurrent interaction of episodic memories: a model of the hippocampal system. *Psychological Review*, *119*, 573.

Leth-Steensen, C., & Marley, A. A. J. (2000). A model of response time effect in symbolic comparison. *Psychological Review, 107,* 62–100.

Lipschutz, S. & Lipson, M. L. (1997). *Theory and Problems of Discrete Mathematics.* McGraw-Hill.

MacLeod, C. M., & Nelson, T. O. (1984). Response latency and response accuracy as measures of memory. *Acta Psychologica, 57*, 215-235.

Moyer, R. S., & Bayer, R. H. (1976). Mental comparison and the symbolic distance effect. *Cognitive Psychology, 8*, 228-246.

Pohl, R., & Schumacher, S. (1991). Handlungssequenzen und Wortlisten als hierarchische lineare Ordnungen: Einflüsse auf den Distanzeffekt [Action sequences and word lists as hierarchical linear orderings: Influences on the distance effect]. *Zeitschrift für experimentelle und angewandte Psychologie, 38,* 43–62.

Potts, G. R. (1972). Information processes used in the encoding of linear orderings. *Journal of Verbal Learning and Verbal Behaviour, 11,* 727-740.

Potts, G. R. (1974). Storing and retrieving information about ordered relationships. *Journal of Experimental Psychology, 103,* 431–439.

R Core Team (2021). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Sadeghi, P., & Arntzen, E. (2018). Eye-movements, training structures, and stimulus equivalence class formation. *The Psychological Record*, *68*, 461-476.

Schubert, T. W. (2005). Your highness: vertical positions as perceptual symbols of power. *Journal of Personality and Social Psychology*, *89*, 1.

Sekuler, R., and Mierkiewicz, D. (1977). Children's judgments of numerical inequality. *Child Development, 48*, 630–633.

Shoben, E. J., Cech, C. G., Schwanenflugel, P. J., & Sailor, K. M. (1989). Serial position effects in comparative judgments. *Journal of Experimental Psychology: Human Perception and Performance, 15*, 273-286.

Singmann, H., Bolker, B., Westfall, J., & Aust, F. (2018). afex: Analysis of Factorial Experiments. R package version 0.20-2. *https://CRAN.R-project.org/package=afex*

Smith, K. H., & Foos, P. W. (1975). Effect of presentation order on the construction of linear orders. *Memory and Cognition, 3,* 614–618.

Smith, K. H., & Mynatt, B. T. (1977). On the time required to construct a simple linear order. *Bulletin of the Psychonomic Society*, *9*(6), 435-438.

Spencer, T. J., & Chase, P. N. (1996). Speed analyses of stimulus equivalence. *Journal of the Experimental Analysis of Behavior, 65*, 643–659.

Trabasso, T. R., & Riley, C. A. (1975). On the construction and use of representations involving linear order. In R. L. Solso (Ed.), *Information processing and cognition: The Loyola Symposium* (pp. 201-229). Hillsdale, NJ: Erlbaum.

Trabasso, T. R., Riley, C. A., & Wilson, E. G. (1975). The representation of linear order and spatial strategies in reasoning: A developmental study. In R. Falmagne (Ed.), *Reasoning: Representation and process* (pp. 201-229). Hillsdale, NJ: Erlbaum.

van Dijck, J. P., Abrahamse, E. L., Majerus, S., & Fias, W. (2013). Spatial attention interacts with serial-order retrieval from verbal working memory. *Psychological Science*, *24*, 1854-1859.

Van Opstal, F., Gevers, W., De Moor, W., and Verguts, T. (2008). Dissecting the symbolic distance effect: Comparison and priming effects in numerical and nonnumerical orders. *Psychonomic Bulletin and Review, 15*, 419–425.

von Hecker, U. (1997). How do logical inference rules help construct social mental models? *Journal of Experimental Social Psychology*, *33*, 367-400.

von Hecker, U., & Klauer, K. C. (2021). Are Rank Orders Mentally Represented by Spatial Arrays? *Frontiers in Psychology, 12*, 613186.

Sedek, G., & Von Hecker, U. (2004). Effects of subclinical depression and aging on generative reasoning about linear orders: Same or different processing limitations?. *Journal of Experimental Psychology: General*, *133*, 237.

von Hecker, U., Klauer, K. C., & Aßfalg, A. (2019). A robust anchoring effect in linear ordering. *Quarterly Journal of Experimental Psychology*, *72*, 2680-2689.

von Hecker, U., Klauer, K. C., Wolf, L., & Fazilat-Pour, M. (2016). Spatial processes in linear ordering. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *42*(7), 1003.

von Hecker, U., Sedek, G., & Brezicka, A. (2013). Impairments in mental model construction and benefits of defocused attention. *European Psychologist*.

Wickelgren, W. A. (1977). Speed-accuracy tradeoff and information processing dynamics. *Acta Psychologica, 41*, 67-85.

Wu, X., & Levy, W. B. (2001). Simulating symbolic distance effects in the transitive inference problem. *Neurocomputing*, *38*, 1603-1610.

Zeithamova, D., & Bowman, C. R. (2020). Generalization and the hippocampus: More than one story? *Neurobiology of Learning and Memory*, *175*, 107317.

Footnotes

1

Note that many other test pairs also present one end element, E, in this case.

2

As one of the reviewers suggested, not only "distance 4" pairs (containing both end points) but also pairs containing only one end point might be susceptible to end-item effects. Whilst we were able to exclude the "two end point" case in Experiment 2 as alternative explanation for the relatively short RTs at distance 4, we are unable, on the basis of the present data, to completely assess the case of "one end point only", as this would require a design including, in an analogous sense, an XABCDEY chain. This is a matter of future research.

3

As one of the reviewers suggested, an alternative explanation for the emergence of a reversed SDE may be that when arranging equivalent elements into a chain, elements of wider distance on the chain might appear less similar to each other than elements positioned close to each other. This possibility might be addressed in future research.

4

Although the main effect of colour condition on accuracy was not significant, Table 3 conveys the impression, as one reviewer suggested, that accuracies in the "colour" condition were somewhat lower than in the "no colour" condition. In an additional analysis, we included all correct AND false responses (that is, trials of all types a, b, and c, see text), now finding that responses in the "colour" condition were overall significantly less accurate than

in the “no colour” condition. It is possible that during learning, participants were not only processing the incoming relations, but simultaneously attempted to form connections between individual persons and colours, thereby incurring an increased workload.

5

As one reviewer had hinted, upon inspecting the accuracy distributions for pair distance = 1, there is reason to suspect a bimodality, predominantly in the accuracy data pertaining to pair distance 1, and predominantly in Experiment 2. We therefore combined the data from all relevant conditions (Experiments 1, 2, and 3 “no colour”) into a global analysis, whereby splitting the sample at the median for pair distance =1 accuracy.

We found that the inversed SDE is clearly present and significant in the subsample with good memory for pair distance = 1 pairs. The distance effect is non-significant in the subsample with below-median memory performance. This strengthens the idea that the reversed SDE emerges as a result of the tendency in participants to retrieve the learned pair distance = 1 elements, and on this basis, in a first attempt, construct the mental model. If, on the other hand, retention of these basic elements is insufficient, there is no reversed SDE because of the floor effect for pair distance = 1.

Table 1. Experiment 1 (combined data from 1a and 1b), Accuracies by pair distance type. Latencies (*ms*). Means and Standard Deviations in brackets, per condition.

	Pair distance type				
	Different Cities	Learning Material (1)	Pair distance (2)	Pair distance (3)	Pair distance (4)
Accuracy	.773	.812	.710	.698	.671
SD	(.179)	(.164)	(.274)	(.285)	(.301)
Latency	1687	1514	1585	1665	1593
SD	(753)	(637)	(889)	(919)	(904)

Table 2. Experiment 2, Accuracies by pair distance type. Latencies (*ms*). Means and Standard Deviations in brackets, per condition.

	Pair distance type				
	Different Cities	Learning Material (1)	Pair distance (2)	Pair distance (3)	Pair distance (4)
Accuracy	.646	.730	.643	.590	.580
SD	(.190)	(.195)	(.219)	(.245)	(.254)
Latency	1717	1481	1518	1599	1522
SD	(693)	(513)	(695)	(853)	(811)

Table 3. Experiment 3, Accuracies by pair distance type, both conditions. Latencies (*ms*). Means and Standard Deviations in brackets, per condition.

	Pair distance type				
	Different Cities	Learning Material (1)	Pair distance (2)	Pair distance (3)	Pair distance (4)
Condition "no colour"					
Accuracy	.860	.911	.919	.884	.844
SD	(.115)	(.082)	(.083)	(.112)	(.147)
Latency	2064	1553	1579	1598	1767
SD	(2595)	(575)	(574)	(681)	(926)
Condition "colour"					
Accuracy	.845	.875	.759	.744	.746
SD	(.199)	(.135)	(.278)	(.298)	(.308)
Latency	2196	1642	1749	1646	1564
SD	(2146)	(538)	(781)	(701)	(687)

Table 4. Experiment 4, Accuracies by pair distance type. Latencies (*ms*). Means and Standard Deviations in brackets, per condition.

	Pair distance type			
	Learning Material (1)	Pair distance (2)	Pair distance (3)	Pair distance (4)
Accuracy	.774	.766	.776	.838
SD	(.121)	(.144)	(.170)	(.159)
Latency	1655	1724	1626	1541
SD	(477)	(637)	(666)	(653)

Appendix A: Modelling of effects

In order to determine which random effect structure to assume, we used in all experiments generalized linear mixed models with random effects for *participants* in case of accuracy data, and linear mixed models with random effects for *participants* in case of latency data. Non-minimal models were compared with the corresponding minimal model for each experiment (see below). If there was a significant difference in fit, the particular type of random slope as specified in the non-minimal model under comparison was then retained for the final model, for the accuracy and latency analyses, respectively, afinal and tfinal. In a second step, these final models were assembled and run in order to evaluate the respective fixed effect structure from those models (see Jaeger, 2008). This strategy thus considers random intercepts and random slopes for the main effects of the experimental design. The analyses employed the statistical programming language R (R Core Team, 2021), using the package lme4 (Bates et al., 2015) and afex (Singmann et al., 2018).

Experiment 1

In the first step, we fitted three models for each data type (a1, a2, a3 for accuracy data, and tm1, tm2, tm3 for latency data). All of these models had the same fixed effect structure, that is, pair distance and block number, as well as their interaction. All models had a random intercept for participants (1 | part). Models a3 and tm3 had only this intercept, so these models are minimal. Models a1 / tm1 also had a random slope for pair distance as function of participant, whereas a2 / tm2 had a random slope for block number instead. These models were then compared using the Chi square difference statistic $\Delta\chi^2$. Models of a given type 1 or 2 were compared with the corresponding model of type 3, the minimal model. If there was a significant difference in fit, the particular type of random slope as specified in the non-minimal model under comparison was then retained for the final model, afinal, resp., tfinal. In a second step, these final models were assembled and run in order to evaluate the respective fixed effect structure from those models (see Jaeger, 2008). This strategy thus considers random intercepts and random slopes for the main effects of the experimental design.

Accuracies

Model	*df*	*AIC*	*BIC*	*loglik*	*deviance*	$\Delta\chi^2$	Δdf	*p*
a3	25	9023.6	9204.0	-4486.8	8973.6			
a1	34	8691.2	8936.6	-4311.6	8623.2	350.36	9	< 2.2e-16 ***
a2	45	8493.3	8818.1	-4201.6	8403.3	570.31	20	< 2.2e-16 ***

afinal = random slopes for pair distance and block number, as a function of participants, are kept.

Latencies

Model	*df*	*AIC*	*BIC*	*loglik*	*deviance*	$\Delta\chi^2$	Δdf	*p*
tm3	26	16578	16756	-8262.9	16526			
tm1	35	16576	16817	-8253.1	16506	19.63	9	0.02034 *
tm2	46	15971	16286	-7939.3	15879	647.25	20	< 2.2e-16 ***

tfinal = random slopes for pair distance and block number, as a function of participants, are kept.

Experiment 2

Model comparisons were performed in the same way as in Experiment 1.

Accuracies

Model	*df*	*AIC*	*BIC*	*loglik*	*deviance*	$\Delta\chi^2$	Δdf	*p*
a3	17	4443.4	4550.5	-2204.7	4409.4			
a1	26	4418.0	4581.8	-2183.0	4366.0	43.45	9	1.782e-06 ***
a2	26	4378.8	4542.7	-2163.4	4326.8	82.586	9	4.943e-14 ***

afinal = random slopes for pair distance and block number, as a function of participants, are kept.

Latencies

Model	*df*	*AIC*	*BIC*	*loglik*	*deviance*	$\Delta\chi^2$	Δdf	*p*
tm3	18	6894.4	6999.5	-3429.2	6858.4			
tm1	27	6904.1	7061.8	-3425.0	6850.1	8.2965	9	0.5046
tm2	27	6836.5	6994.3	-3391.3	6782.5	75.83	9	1.083e-12

tfinal = random slopes for block number, as a function of participants, are kept.

Experiment 3

In the first step, we fitted three models for each data type (a1, a2, a3 for accuracy data, and tm1, tm2, tm3 for latency data). All of these models had the same fixed effect structure, that is, colour version, pair distance and block number, as well as their interactions. All models

had a random intercept for participants (1 | part). Models a3 and tm3 had only this intercept, so these models are minimal. Models a1 / tm1 also had a random slope for pair distance as function of participant, whereas a2 / tm2 had a random slope for block number instead. These models were then compared using the Chi square difference statistic $\Delta\chi^2$. Models of a given type 1 or 2 were compared with the corresponding model of type 3, the minimal model. If there was a significant difference in fit, the particular type of random slope as specified in the non-minimal model under comparison was then retained for the final model, afinal, resp., tfinal. In a second step, these final models were assembled and run in order to evaluate the respective fixed effect structure from those models (see Jaeger, 2008). This strategy thus considers random intercepts and random slopes for the main effects of the experimental design.

Accuracies

Model	*df*	*AIC*	*BIC*	*loglik*	*deviance*	$\Delta\chi^2$	Δdf	*p*
a3	9	4055.8	4116.6	-2018.9	4037.8			
a1	58	4010.2	4402.0	-1947.1	3894.2	143.58	49	3.221e-11 ***
a2	69	3841.9	4308.0	-1852.0	3703.9	333.87	60	< 2.2e-16 ***

afinal = random slopes for pair distance and block number, as a function of participants, are kept.

Latencies

Model	*df*	*AIC*	*BIC*	*loglik*	*deviance*	$\Delta\chi^2$	Δdf	*p*
tm3	50	10876	11202	-5387.8	10776			
tm1	59	10867	11253	-5374.6	10749	26.401	9	0.001756 **
tm2	70	10785	11242	-5322.4	10645	130.83	20	< 2.2e-16 ***

tfinal = random slopes for pair distance and block number, as a function of participants, are kept.

Experiment 4

Model comparisons were performed in the same way as in Experiment 1 and 2.

Accuracies

Model	*df*	*AIC*	*BIC*	*loglik*	*deviance*	$\Delta\chi^2$	Δdf	*p*
a3	25	11324	11508	-5636.9	11274			
a1	34	11225	11476	-5578.6	11157	116.74	9	< 2.2e-16 ***
a2	45	10567	10899	-5238.6	10477	796.64	20	< 2.2e-16 ***

afinal = random slopes for pair distance and block number, as a function of participants, are kept.

Latencies

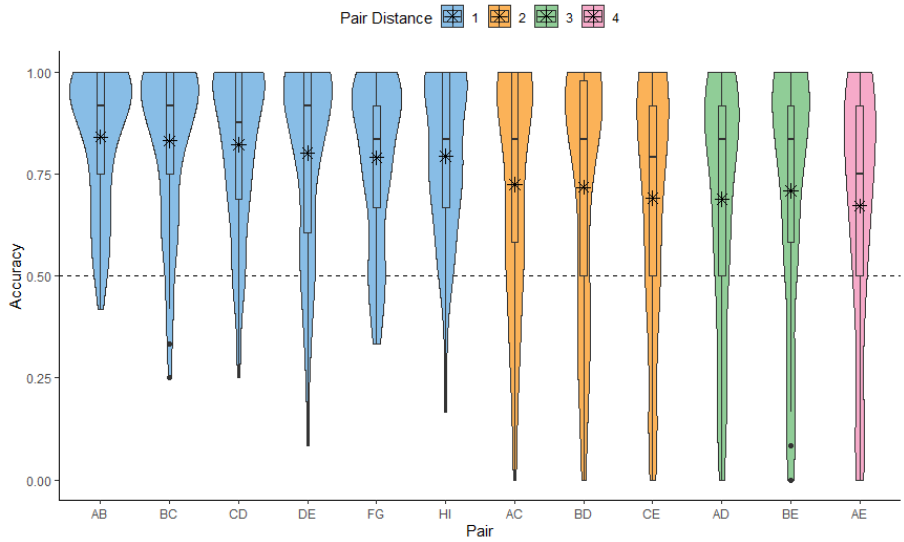
Model	*df*	*AIC*	*BIC*	*loglik*	*deviance*	$\Delta\chi^2$	Δdf	*p*
tm3	26	19747	19931	-9847.5	19695			
tm1	35	19653	19901	-9791.5	19583	112.03	9	< 2.2e-16 ***
tm2	46	19508	19833	-9707.8	19416	279.42	20	< 2.2e-16 ***

tfinal = random slopes for pair distance and block number, as a function of participants, are kept.

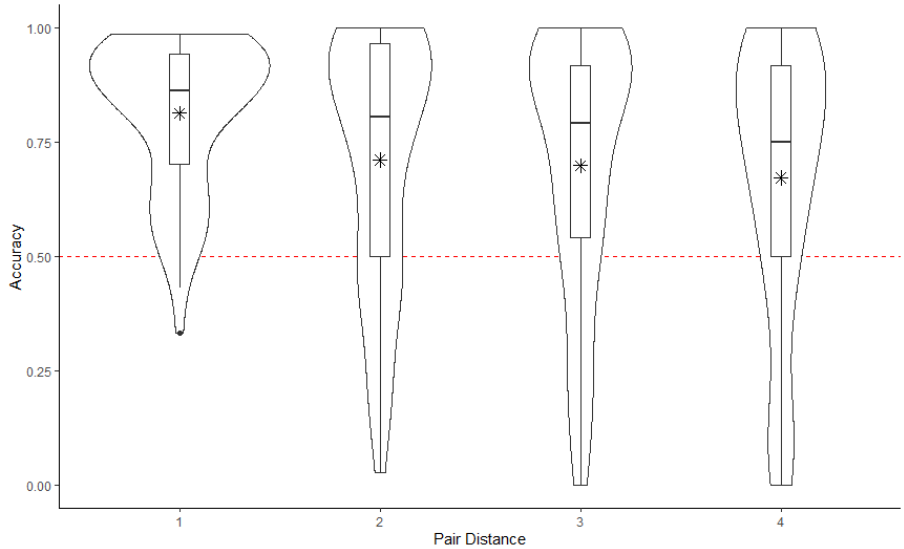
Appendix B: Violin Plots

The plots are arranged in the following way. For each Experiment (E1, E2, E3, and E4) there are separate plots showing accuracy data and reaction time data. In each case, these data are presented in two versions: one (coloured) plot showing distributions for all individual relations separately. In addition, one (black and white) plot showing groups of relations summarised according to pair distance. In case of E3, there are separate sets of plots for the "white only" condition (cv0) and for the "coloured" condition (cv1).

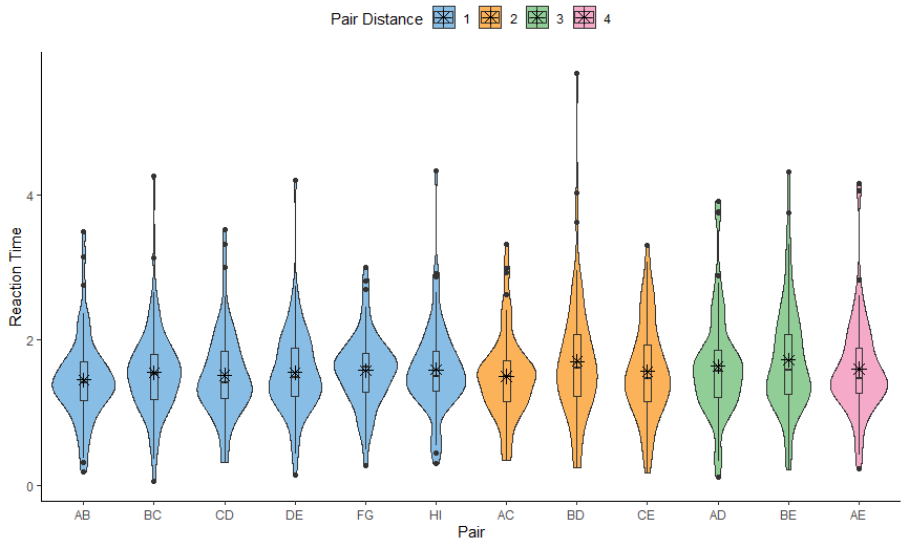
Experiment 1, accuracy data, individual relations



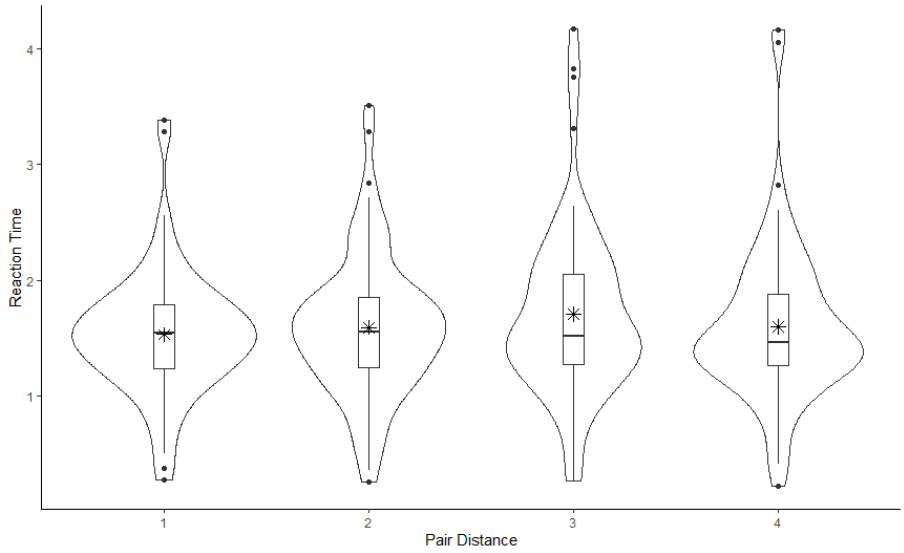
Experiment 1, accuracy data, pooled by pair distance



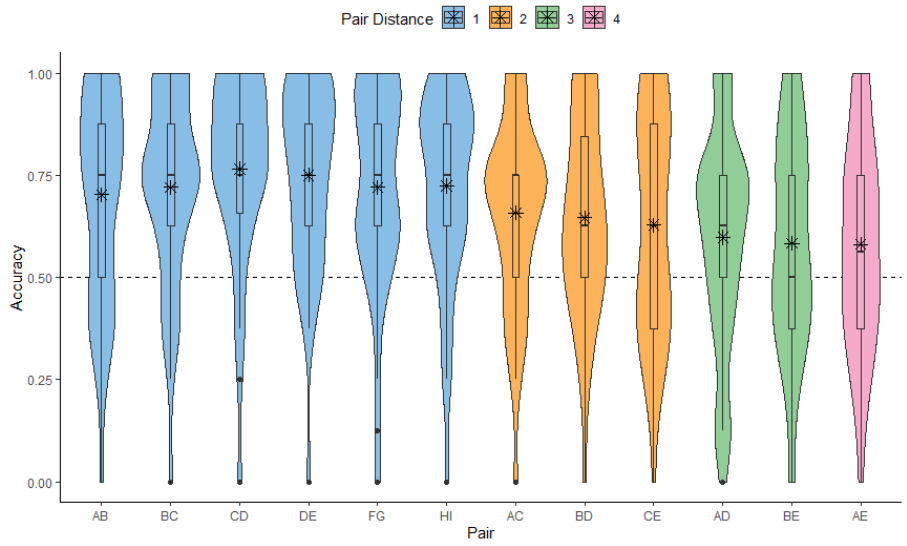
Experiment 1, reaction time data, individual relations



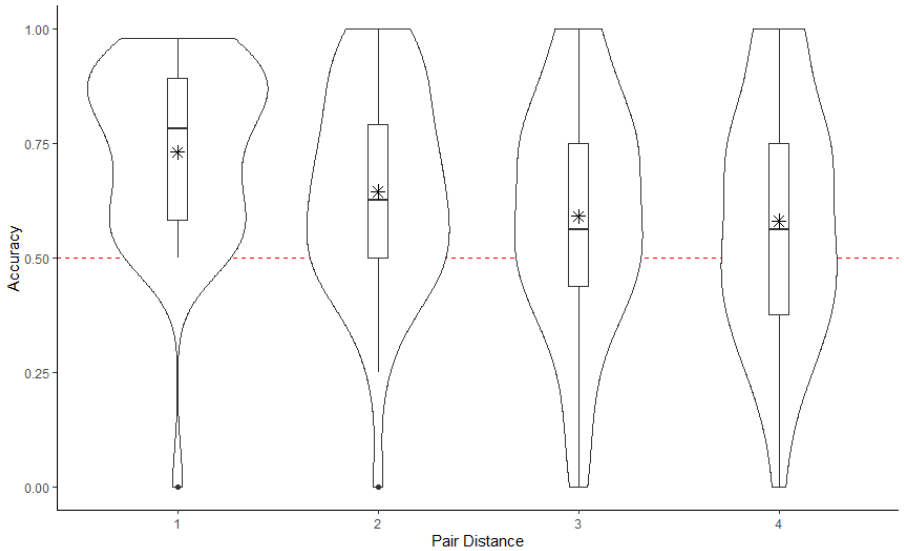
Experiment 1, reaction time data, pooled by pair distance



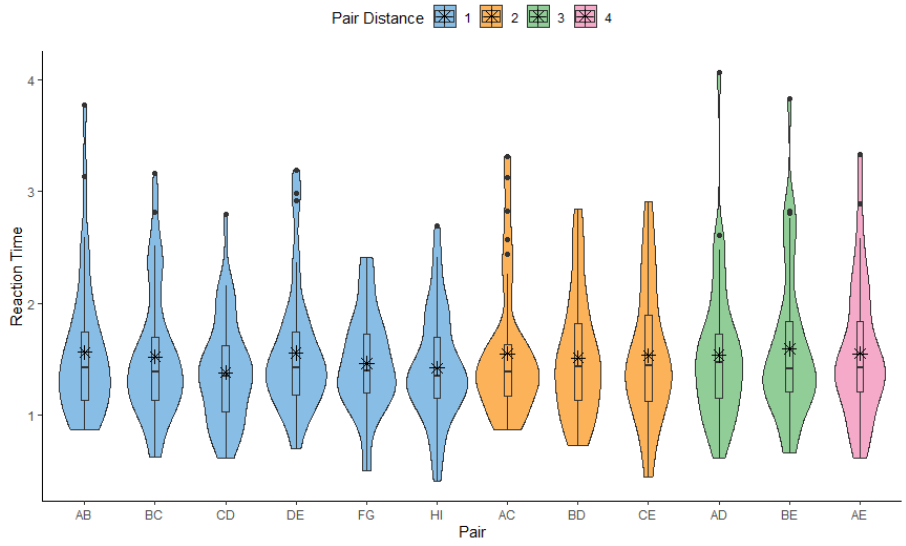
Experiment 2, accuracy data, individual relations



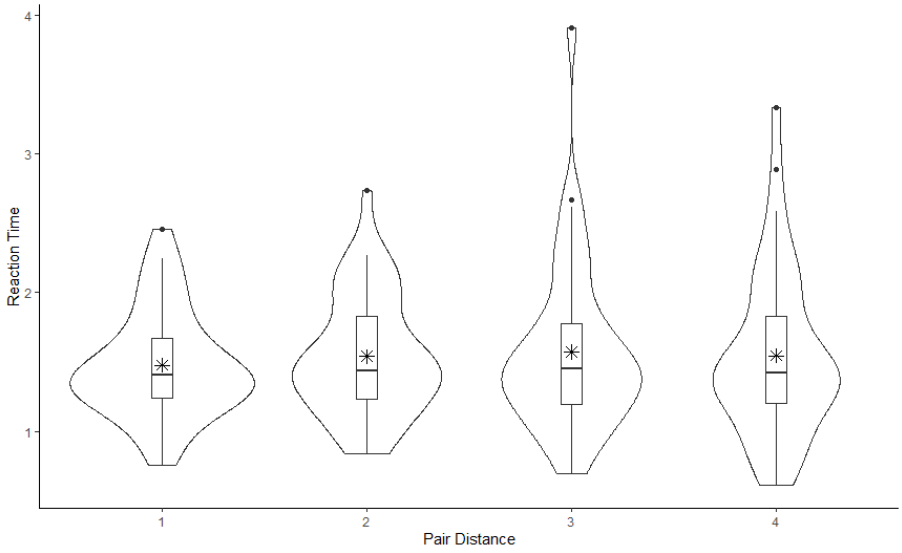
Experiment 2, accuracy data, pooled by pair distance



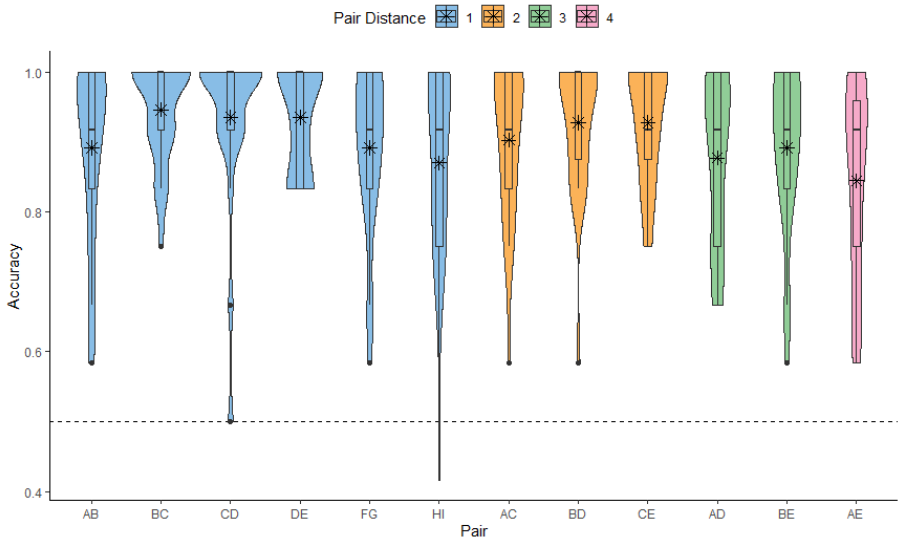
Experiment 2, reaction time data, individual relations



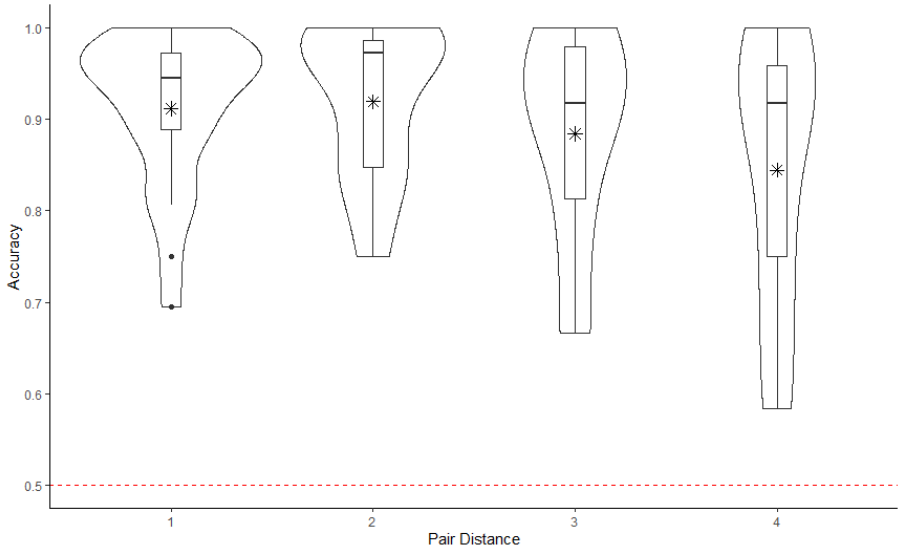
Experiment 2, reaction time data, pooled by pair distance



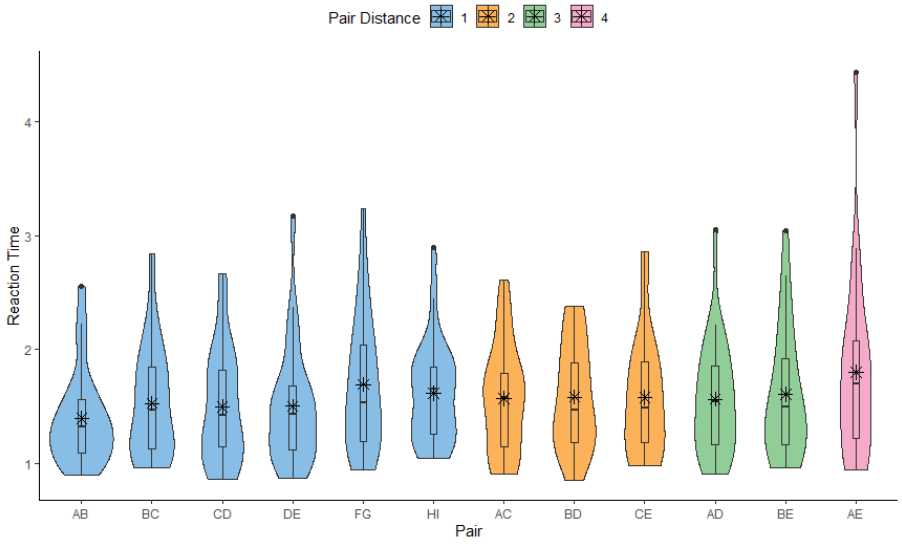
Experiment 3, “no colour” condition, accuracy data, individual relations



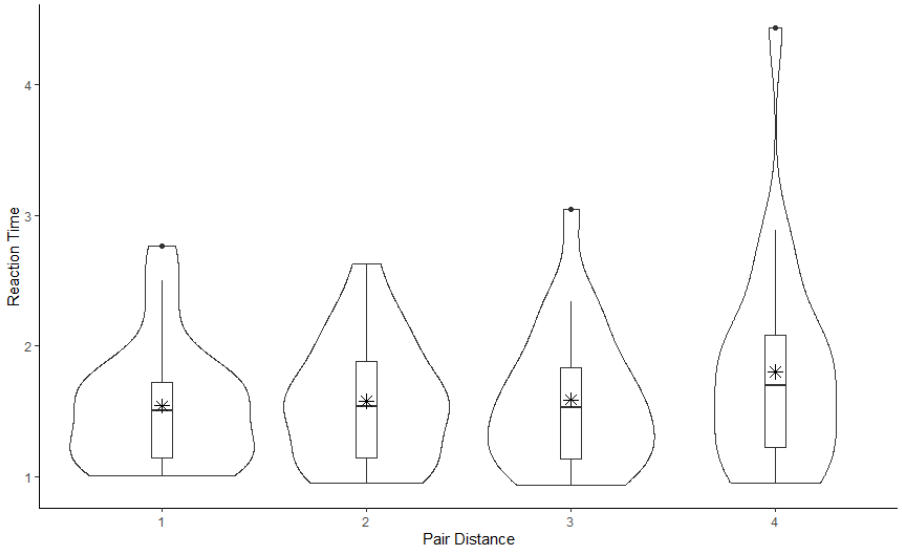
Experiment 3, “no colour” condition, accuracy data, pooled by pair distance



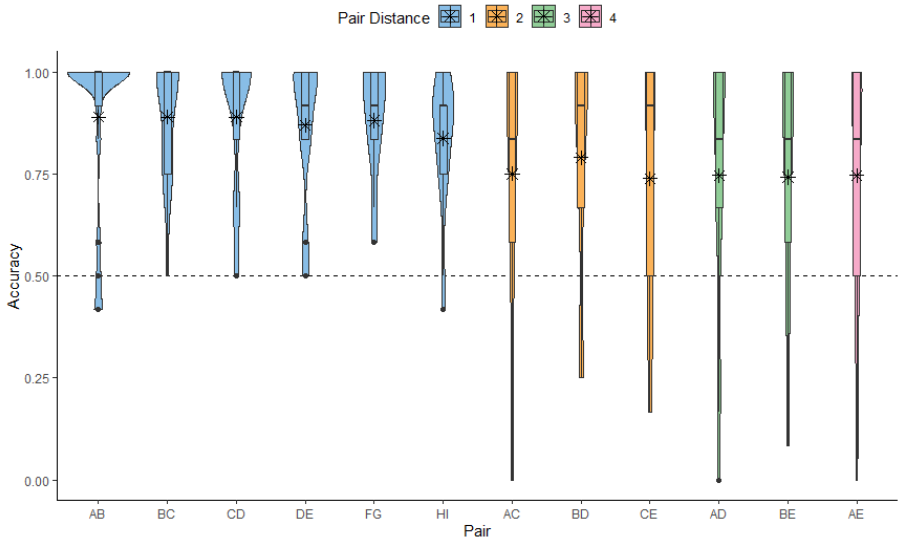
Experiment 3, “no colour” condition, reaction time data, individual relations



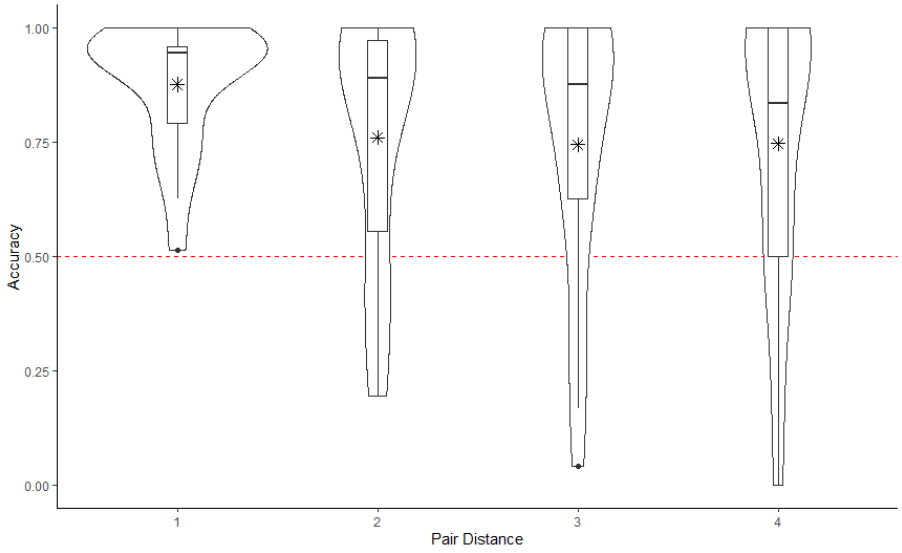
Experiment 3, “no colour” condition, reaction time data, pooled by pair distance



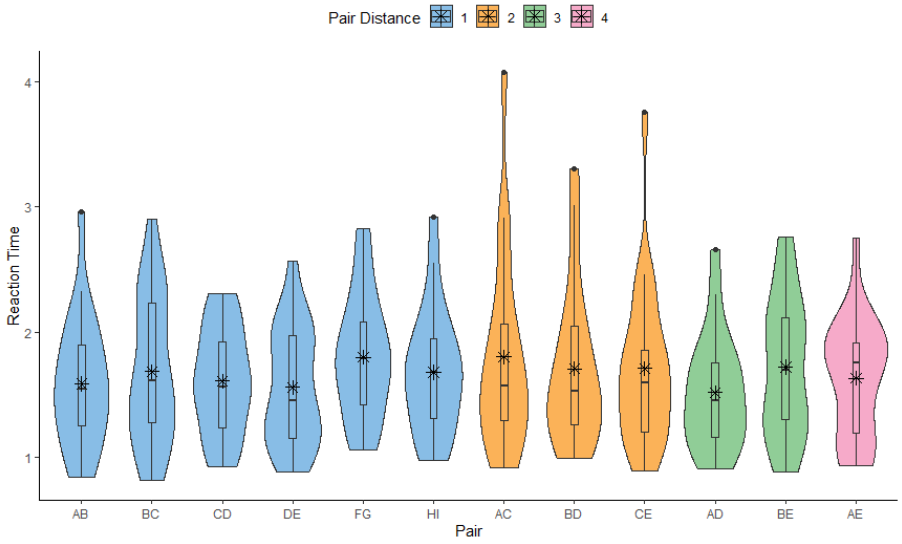
Experiment 3, “colour” condition, accuracy data, individual relations



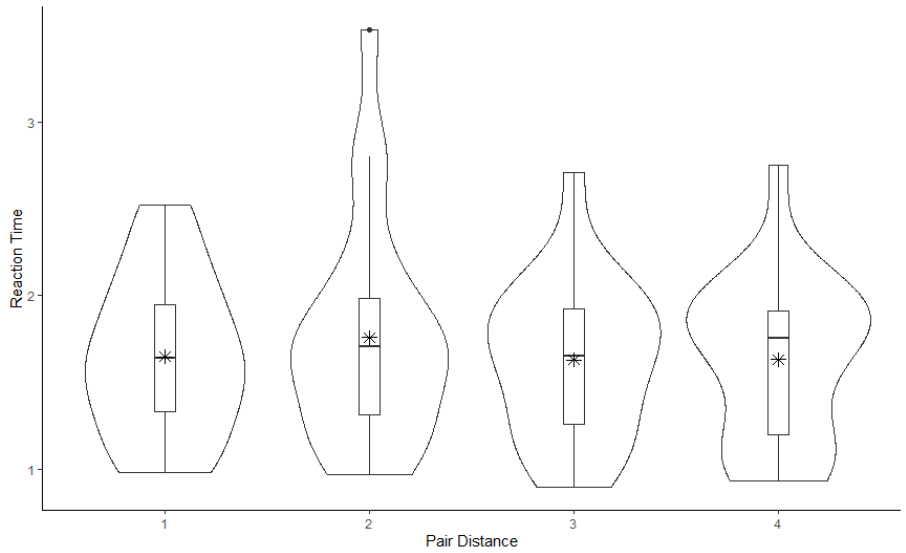
Experiment 3, “colour” condition, accuracy data, pooled by pair distance



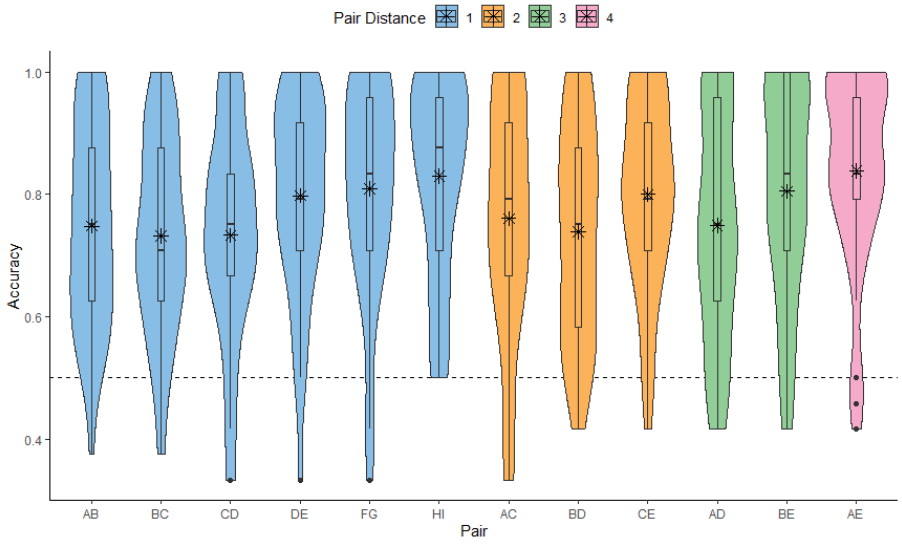
Experiment 3, “colour” condition, reaction time data, individual relations



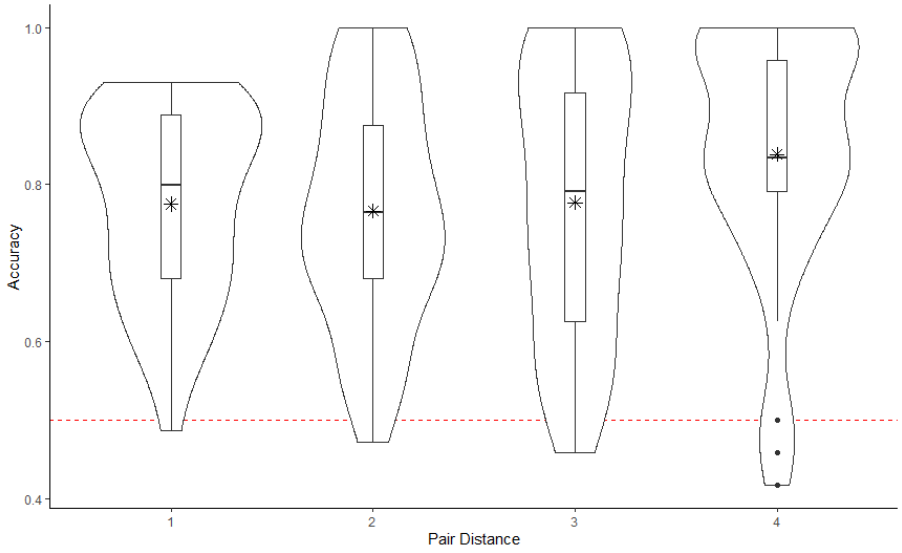
Experiment 3, “colour” condition, reaction time data, pooled by pair distance



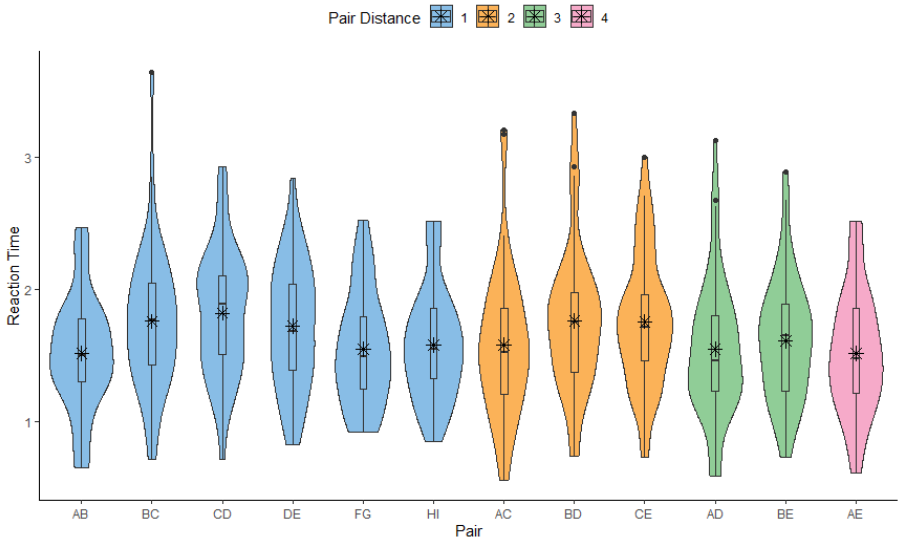
Experiment 4, accuracy data, individual relations



Experiment 4, accuracy data, pooled by pair distance



Experiment 4, reaction time data, individual relations



Experiment 4, reaction time data, pooled by pair distance