Evidence accumulation in same-different judgments: Integrating featural similarity with structural knowledge using a linear ballistic accumulator

Andrew T. Hendrickson
Cognitive Science & Artificial Intelligence Department
Tilburg University

Daniel J. Navarro
School of Psychology
University of New South Wales

Chris Donkin
School of Psychology
University of New South Wales

Abstract

Stimulus similarity plays a fundamental role in human cognition, shaping theoretical accounts of category learning, inductive reasoning, memory, and others. In this paper we introduce an evidence accumulation model for similarity based decisions that successfully accounts for the complete joint distribution over choice and response time across different stimuli, and how these distributions change systematically as a function of instructional demands. The modeling framework captures the way in which information about simple feature matches drives the early stages of stimulus processing, whereas the later stages are more heavily influenced by structural knowledge about the stimulus. Despite recent work showing that single process models are very often able to accommodate phenomena that ostensibly provide evidence for multiple processes, we show that no single process model provides a qualitatively reasonable fit to the results from two experiments.

Introduction

Similarity plays an important role in cognitive science. Theories of categorization, induction and memory are all substantially reliant on some notion of stimulus similarity. Empirically, measures of stimulus similarity are used to supply mental representations via statistical techniques like multidimensional scaling, additive clustering and others. Yet similarity itself is notoriously slippery. It cannot be defined logically, on the basis of shared properties (Goodman, 1972), and it is sensitive to a variety of factors that suggest that the sense of similarity is not a primitive, but rather is constructed on the fly by the cognitive

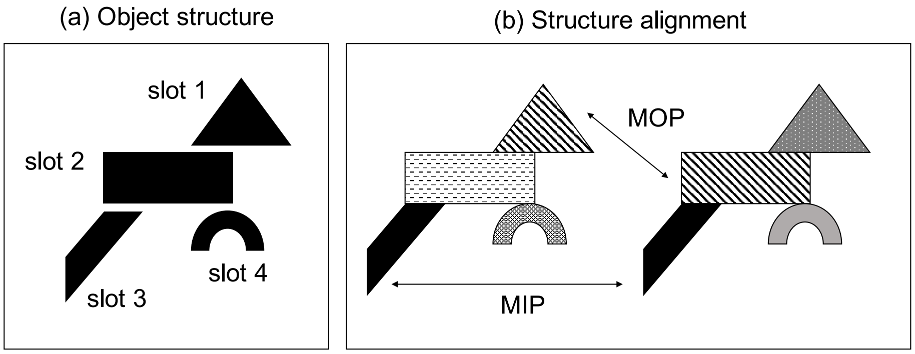

Figure 1. The role of structure when assessing similarity. Panel a depicts a simple object composed of four distinct shapes or "slots" that can take on different feature "values" (color, texture, etc). When assessing the similarity between two such objects (panel b) there is meaningful distinction to be made between feature matches in place (MIPs), when the same slot takes on the same value in different objects, and feature matches out of place (MOPs), in which the same feature appears in two objects, but appears in different slots. In general, MIPs make a greater contribution to similarity than MOPs.

system (Medin, Goldstone, & Gentner, 1993). The complexity of judgments based on similarity is evidenced in a number of ways. People use different information to judge similarity depending on context (Goldstone, Medin, & Halberstadt, 1997), on whether similarity or dissimilarity is to be judged (Medin, Goldstone, & Gentner, 1990), the direction on which comparisons are to be made (Tversky, 1977), and a variety of other factors. The relationship between perceptual similarity and conceptual knowledge is not straightforward (Smith & Heise, 1992), and it is not uncommon for similarity judgments to reflect both.

Acknowledging the many facets of similarity opens up the question of time-dependent processing. To the extent that different kinds of information take longer to retrieve from memory or require complicated processing to extract, different information may be accessible to people at different time points in processing. From this perspective, our theories must not merely describe the information people use to judge similarity: we must also consider *when* different sources of information become available to the decision maker. There are many approaches to similarity that can accommodate how different information might be used in different *contexts* (e.g., shifts in attention, feature weighting, etc), but fewer that consider how different information may be available during the time course of a *single* decision: historically, the view of similarity as a perceptual primitive discouraged systematic investigation of the time course of information processing (see Medin et al., 1993). Nevertheless, there are many possible mechanisms that might play a role in determining the information available at different points in a similarity comparison: stimulus features might vary in their salience, in the amount of cognitive processing required to activate them, and in the ease by which they can be cued and recalled from memory. Any of these factors could influence when these features are available and can contribute to similarity. To the extent that they do, the perceived similarity between two items should systematically change during the time course of a decision.

Fast features and slow structure

One particularly important factor to consider is how the relative contribution of simple featural information and more complex relational knowledge changes during the time course of similarity comparison. There are reasons to expect systematic changes over time. For instance, in the vision literature it is well established that simple featural information can be detected very rapidly but slower, attention-dependent processing is required for raw features to be bound together into coherent object representations (Treisman & Gelade, 1980), producing illusory conjunctions when object features are not properly aligned together (Treisman & Schmidt, 1982). This binding process is critical for forming structured stimulus representations because it allows object representations to specify not only the set of perceptual features that an item has, but the relationships among them. A variation of this idea is mirrored in "structural alignment" approaches to similarity comparison (Goldstone & Medin, 1994; Goldstone, 1994). Inspired by theories of analogical reasoning (Falkenhainer, Forbus, & Gentner, 1989; Gentner, 1983, 1989; Hummel & Holyoak, 1997), these models assume that featural information is available early in the comparison process, but the mapping from one stimulus to another is a slower process because it relies on the construction and comparison of structured object representations. Accordingly, it is reasonable to propose that simple feature match information becomes available for similarity comparison before structural information about the relationships between features.

To illustrate the different ways in which structural and featural information can contribute to similarity, consider the simplified framework depicted in Figure 1. When two objects share a feature (e.g., black color) that appears in the same context (e.g., leg), this commonality is referred to as a *match in place* (MIP). When the feature appears (e.g., stripes) in different contexts (e.g., head vs body), this commonality is referred to as a *match out of place*. Drawing on theories of analogy and metaphor (Falkenhainer et al., 1989; Gentner, 1983, 1989), it has been argued that the perception of similarity is highly sensitive to this kind of relational knowledge: a MIP contributes more to perceived similarity than a MOP (Goldstone & Medin, 1994). However, as noted above, this theoretical perspective also makes the claim that the relational information that allows the colors of one stimulus to be matched against colors in another stimulus arrives more slowly, because the cognitive system is engaged in an active process of aligning the representations of one stimulus against the other. At a minimum, the distinction between MIPs and MOPs only makes sense *after* features are bound to contexts or slots: prior to the resolution of this binding process, it should not be possible to rely on this knowledge, and as a consequence structural alignment models predict that perceived similarity changes over time: early in processing people should not make a distinction between MIPs and MOPs, and only later do MIPs become more important.

The idea that there are multiple distinct sources of evidence (simple feature matches and relational knowledge) that contribute to perceptual similarity is an appealing one, and one that appears to be supported by evidence. Typical results (Goldstone & Medin, 1994) rely upon response deadline tasks, which force participants to respond after a fixed amount of time. In these designs, an interaction effect between choice probabilities and response deadline is interpreted as the signature of time-dependent similarity. The logic for this is sensible: if people produce a qualitatively different pattern of responding at time 2 than they do if processing is stopped at time 1, it seems reasonable to conclude that qualitatively different evidence is available at the two points in time.

The perils of proposing multiple processes

As reasonable as the preceding discussion sounds, recent research suggests that considerable caution is required. Theoretical work has shown that it is much harder to find diagnostic tests of single-process versus dual-processes accounts than is typically acknowledged (Newell & Dunn, 2008). Indeed, in recent years dual-process accounts of many different phenomena have struggled to outperform their single-process counterparts in recognition memory (Dunn, 2008, 2004), reasoning (Lassiter & Goodman, 2015; Stephens, Dunn, & Hayes, in press), and categorization (Newell, Dunn, & Kalish, 2011), even with respect to the original phenomena used to justify the development of such accounts. Moreover, when considering the evidence highlighting the tight links between ostensibly different inference problems such as similarity, identification, categorization, and recognition (Nosofsky, 1986; Hawkins, Hayes, & Heit, 2016) it would seem prudent to be very cautious when proposing that two qualitatively different processes are involved in making a single kind of judgment.

In the case of similarity judgments produced under time constraints, this concern seems especially pertinent. As the literature on choice response times indicates (Luce, 1986; Ratcliff, 1978) the interpretation of deadline tasks or signal to respond tasks can be quite difficult (Ratcliff, 2006). More generally, significant interaction effects in an ANOVA need not be indicative of any fundamental change in processing, merely shifts in decision criteria (Dube, Rotello, & Heit, 2010). It is entirely possible that these interaction effects are merely artifacts of strategic behavior and do not reflect true differences in the underlying similarity perceived by participants.

To illustrate the concern, consider the finding by Goldstone and Medin (1994), namely that the number of MOPs has the greatest influence on "same vs different" judgments when deadlines are short, and the number of MIPs has the greatest influence when deadlines are long. On the surface this result suggests that the relative *weight* of MIPs and MOPs is changing over time. However, this need not be so. To foreshadow the computational modeling later in the paper, consider the simple evidence accumulation depicted in Figure 2a, based the linear ballistic accumulator framework for simple choice (LBA; Brown & Heathcote, 2008). In this model, we assume there exists a single underlying similarity signal that does not change over time, one that weighs MIPs more heavily than MOPs. This signal drives an evidence accumulation process that has three possible outcomes: it produces a "same" response if a similarity-driven accumulator reaches threshold before a dissimilarity-driven accumulator does, produces a "different" response if the dissimilarity process terminates first, and produces a random answer if the deadline arrives before either process terminates.[1] The critical thing to note that in this model, the relative contribution of MIPs and MOPs in driving the evidence accumulation process does not change over time, and as such is not consistent with the hypothesis of time-dependent changes in similarity comparison. Nevertheless, as shown in Figure 2b, the model reproduces the same qualitative pattern of

[1] Specifically, the model simulations in Figure 2b rely on a standard LBA model in which the mean drift rate is .09 times the number of MOPs (left panel) or .18 times the number of MIPs (right), the trial-to-trial drift variability is .2, and the response threshold is 20. If the decision time exceeds 20 (short) or 60 (long), a random answer is given.

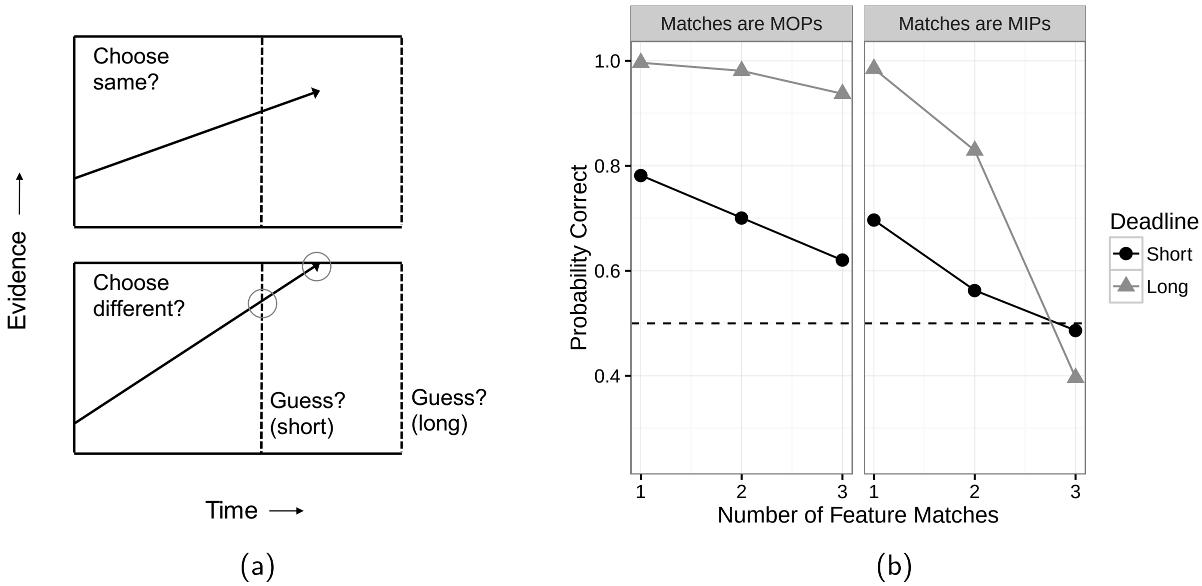

Figure 2. Interaction effects in response deadline tasks can occur even when subjective stimulus similarity does not change during processing. A simple linear ballistic evidence accumulation model with drift rates determined by a linear combination of MIPs and MOPs (panel a) can produce the empirically-observed deadline interaction effect, in which MIPs have their largest effect on choices at long deadlines and MOPs have their largest effect at short deadlines (panel b). See text for details.

responses found by Goldstone and Medin (1994): the number of MOPs has a stronger effect on accuracy under short deadlines than under long deadlines, whereas the number of MIPs more strongly influences later responses. In short, a significant interaction in a response deadline task does not constitute strong evidence in favor of time-dependent changes in perceived similarity.

Summary

The model simulations in Figure 2 raise a worrying possibility for theories advocating more than one signal driving similarity: perhaps the pattern of results observed in deadline tasks is caused solely by the fact that MIPs are more salient than MOPs, and do not reflect any change over time in how similarity is computed. With that in mind, our goal in this paper is twofold. First, we present experimental evidence from two reaction time tasks based on Goldstone and Medin (1994) that do not rely on experimenter imposed deadlines, instead relying on an instructional manipulation that emphasizes speed or accuracy to induce subjects to adjust their own decision making criterion. This approach is standard in the choice reaction time literature (Luce, 1986; Ratcliff & Smith, 2004) and avoids many of the difficulties associated with deadline tasks. Second, we develop a family of evidence accumulation models for same-different judgments that allows us to explicitly test models that treat feature match information and structural information as distinct sources of evidence (and hence can have differential impact at different time points) against models

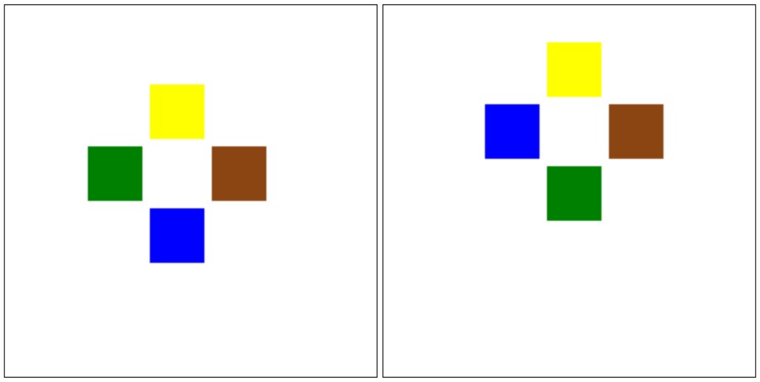

Figure 3. A sample trial. The stimulus display is divided into two frames, with one object per frame. Each object consists of four colored squares arranged into a cross shape, and the task is to indicate whether the objects are identical. These stimuli are characterized by two MIPs (yellow and brown squares in identical positions) and two MOPs (the green and blue squares are in different positions), and thus DIFFERENT stimulus items that belong to the `2[2]` condition.

that assume that similarity is a "single" time-homogeneous construct that – while potentially weighting MIPs and MOPs differently – requires the relative importance of any one source to remain constant over time. Despite trying several possibilities, we find that no such "single process" account of similarity can accommodate our data in an even remotely plausible fashion, whereas a simple dual process model that separates feature match information from structural alignment information accounts for the data naturally.

Experiment 1

Our experimental work adapts Experiment 2 from Goldstone and Medin (1994) who investigated how people assess the similarity between structured stimuli. Consider the stimuli shown in Figure 3. Each "cross" stimulus consists of four "slots" each of which can be occupied by squares of different color. The spatial arrangement of the slots ensures that the stimuli are not merely four blobs of color: instead, they are organized into a specific configuration, ensuring that the two items shown in Figure 3 are distinct. Participants were presented with pairs of stimuli and asked to judge if they were identical or not (i.e., a same-different task). If the same color appears in the same position for both stimuli, that feature is a MIP, whereas if it appears in a different position it is a MOP. A color that appears in one stimulus but not the other is a mismatch. In their original work Goldstone and Medin (1994) found that the relative importance of MIPs and MOPs appeared to change as a function of response deadline, but as discussed earlier it is difficult to know how to interpret this finding. In our approach we remove the response deadlines, replacing it with an instructional manipulation encouraging people to emphasize *speed* or *accuracy* in making their decisions.

Condition	Stimulus A	Stimulus B	Count
0[0]	ABCD	EFGH	50
0[2]	ABCD	BAEF	50
2[0]	ABCD	ABEF	50
0[4]	ABCD	BCDA	50
2[2]	ABCD	ABDC	50
SAME: 4[0]	ABCD	ABCD	250

Table 1
The six stimulus types in Experiment 1. The letters A through H indicate unique randomly chosen colors, and each column corresponds to one of the four positions in the cross. The condition names reflect the number of MIPs and MOPs in each stimulus respectively (e.g. the `0[2]` condition includes stimuli that have 0 MIPs and 2 MOPs), and the count column indicates the number of presentations of each stimulus type.

Method

Participants. 250 people were recruited from Amazon's Mechanical Turk and paid US$1.50 for approximately 18 minutes of work. Data was collected from 262 participants. 13 were excluded from all analyses for failing to complete the task. Of the remaining 249, 43% were female. Ages ranged from 18 to 66 years (mean: 34). 84% were from the USA, 14% from India, and 2% from other countries.

Materials & procedure. The stimuli used in the task are illustrated in Figure 3. Each colored square was 40 pixels wide, and the positions of the crosses randomly jittered slightly from trial to trial within each frame. 8 colors were used (red, blue, green, yellow, turquoise, brown, gray, and orange). Colors were assigned randomly subject to the constraints outlined in Table 1, which lists 6 logically distinct stimulus types. On half of the 500 trials the two items were identical (i.e., consisted of 4 MIPs). On the other half of trials, at least two colors did not match. Table 1 outlines the five stimulus types that span all possible patterns of mismatches between the two items. The `0[0]` stimuli, for instance, share no color features, whereas in the `0[4]` the items share the same four colors, but no color appeared in the same position. In potentially the most confusing case (the `2[2]` condition) the same four colors appear in both items, two in the same positions of both items and two with their positions swapped.

The "cross" stimuli and the same-different task were explained to all participants and participants were required to correctly answer four instruction check questions before proceeding to the experiment. For the 116 participants randomly assigned to the *speed* condition, the instructions asked them to respond as quickly as possible. During the task itself, if they did not respond within 1000 ms then after the trial they were informed that response was "too slow." The remaining 133 participants were assigned to the *accuracy* condition and were encouraged to answer as accurately as possible, and were provided with feedback indicating their response was "incorrect" after incorrect responses. Feedback was presented for 700 ms. Every 50 trials participants were given a break and told their accuracy and average response time from the previous block. During these breaks between blocks participants were reminded to be fast or accurate depending on their condition.

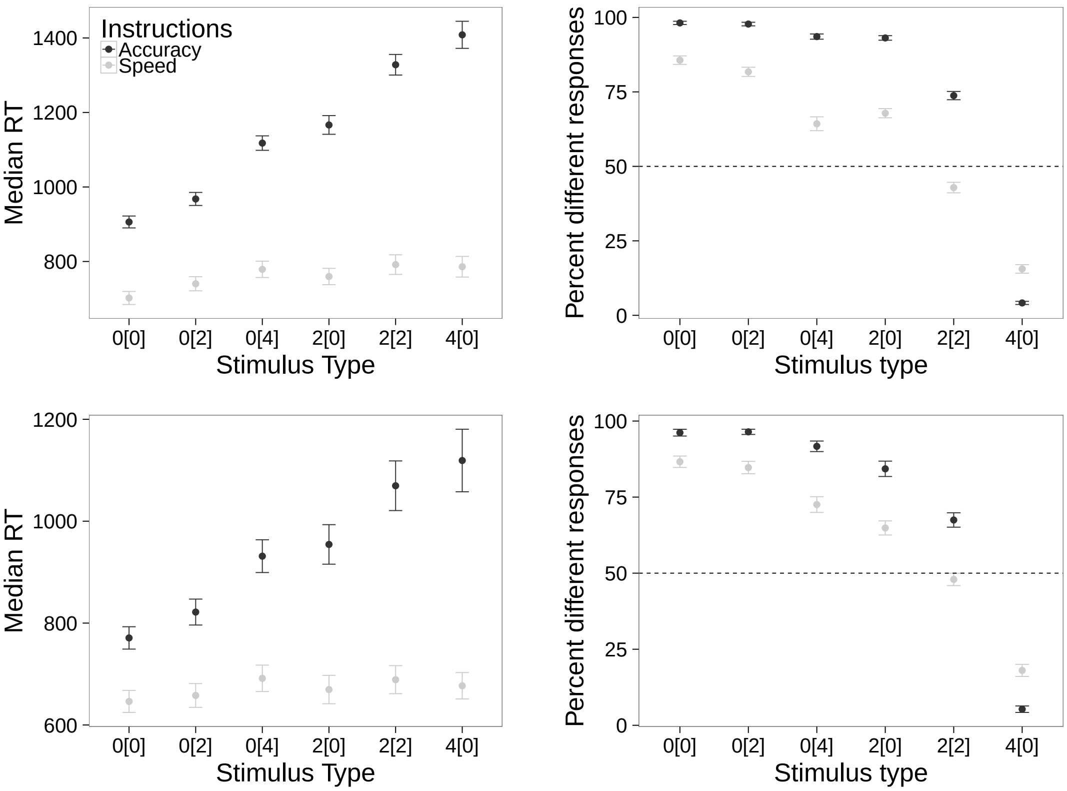

Figure 4. Median response times (left column) and accuracy (right column) for participants in Experiment 1 (top row) and Experiment 2 (bottom row). Data are plotted separately for each stimulus condition, with data from the *Speed* condition are plotted as gray points and data from the *Accuracy* condition are plotted in black. The pattern of results in both experiments is unambiguous: responses are slower in the *Accuracy* condition and RTs are systematically influenced by stimulus type, whereas response times for the *Speed* condition are relatively constant. Under *Accuracy* instructions people make more correct decisions, but both instructions sets reveal an effect of stimulus type. Error bars correspond to 1 standard error and dashed lines indicate chance performance.

Results

The left column of Figure 4 shows the median response times (RT) and the right column shows the corresponding proportion of 'different' responses. Visual inspection of the plot suggests substantial and systematic differences in choices and RT across the stimulus types and instruction conditions. When given *speed* instructions, participants show almost no differences in response time across conditions. However, as the two stimuli become increasingly similar – as we move from left to right in the figures, we see that participants become increasingly less likely to respond 'different'. Most remarkably, we see that participants in the *speed* condition were more likely to endorse the `2[2]` stimulus type as 'same' than 'different'. To foreshadow our conclusions, this result will prove very difficult for any model to accommodate without including a mechanism that allows the relative importance of MIPs and MOPs to shift over time.

Under *accuracy* conditions, we see that participants are indeed much more accurate. Even in the most difficult condition, participants maintain an accuracy of around 75%. In contrast to the *speed* condition, we now see a gradual increase in response time as the pair of stimuli become more similar. Responses are slowest when the two items are identical, but are also relatively slow in the difficult `2[2]` condition. So, unlike the speed-emphasis condition, participants do not reliably respond 'same' to the `2[2]` stimuli. However, this accuracy seems to come at a substantial cost to response time.

A Bayesian data analysis confirms these observations. For the accuracy data, a Bayesian mixed-effects ANOVA (see Rouder, Morey, Speckman, & Province, 2012) shows an effect of instruction condition (BF $> 10^{31}$) and an effect of stimulus type (BF $> 10^{185}$).[2] The full model containing instruction condition, stimulus type and an interaction between the two factors had the largest Bayes factor (BF $> 10^{245}$). A similar pattern emerges when analyzing median reaction times. A Bayesian mixed-effects ANOVA shows an effect of instruction condition (BF $> 10^{25}$) and an effect of stimulus type (BF $> 10^{112}$). As before, the full model containing instruction condition, stimulus type and an interaction between the two factors had the largest Bayes factor (BF $> 10^{238}$).

Discussion

The results from Experiment 1 are qualitatively consistent with the original findings by Goldstone and Medin (1994). Where they found a significant interaction between stimulus type and the response deadline for both accuracy and response time, we found reliable interactions between stimulus type and the instruction condition for both measures. The fact that we observe these effects without an experimenter imposed response deadline is reassuring, and suggests that the original findings are not a methodological artifact. Moreover, the pattern of results in our data is itself intriguing: the trade-off between response time and accuracy is complex, with no effect of similarity on RT under *speed* instructions, and increasing RT with similarity under *accuracy* instructions. The accuracy data is equally interesting, with performance on the `2[2]` stimuli (Figure 4) being below chance for the *speed* instructions and above chance for the *accuracy* instructions. Intuitively, this pattern of results seems to place strong constraints on models of similarity-based decisions.

Experiment 2

The between-subject nature of the speed-accuracy instruction manipulation in Experiment 1 makes it difficult to properly estimate a set of parameters for each participant to compare the prediction of computational models. We address this issue in Experiment 2 by replicating this experiment in a within-subject design.

Method

Participants. 50 workers were recruited from Amazon's Mechanical Turk and paid US$3.25 for approximately 28 minutes of work. During recruitment, 5 workers failed to complete all 1000 trials and were replaced in the sample. 46% were female and they ranged in age from 21 to 63 years (mean: 34). All participants reported being from the USA.

[2] All analyses were conducted using the Bayes factor package in R (Morey & Rouder, 2015).

Materials and procedure. Experiment 2 was a direct replication of Experiment 1, except that the instructions stressing speed or accuracy became a within-subject manipulation. The order of instructions were randomized for each participant. Each participant completed a total of 1000 trials.

Results and discussion

The results of Experiment 2 replicate the findings from Experiment 1 for both response time and accuracy (Figure 4). Bayesian mixed-effects ANOVA for accuracy shows an effect of instruction condition (BF$> 10^{22}$) and an effect of stimulus type (BF $> 10^{61}$). The full model containing instruction condition, stimulus type and an interaction between the two factors had the largest Bayes factor (BF $> 10^{105}$). A similar pattern emerges from the ANOVA on response time. We find an effect of instruction condition (BF $> 10^{53}$) and an effect of stimulus type (BF $> 10^{7}$). As before, the full model containing instruction condition, stimulus type and an interaction between the two factors had the largest Bayes factor (BF $> 10^{78}$).

Evidence accumulation models of structured similarity

The accuracy and median response time data suggest that the observer may utilize feature information more than relational information earlier in the decision process. However, it would be premature to draw any strong conclusions about such factors based on the empirical data alone. We now show that a comprehensive set of single-process similarity models qualitatively fail to capture the observed data. All of these single-process models assume that a single, stationary, stream of information about the similarity between the two stimuli is accumulated over the course of a trial. We then show that a model allowing for different sources of information – one feature-based and one structural – provides an impressive account of our data.

We use the linear ballistic accumulator (LBA) model to formally characterize the decision making process of participants in Experiment 2 (Brown & Heathcote, 2008). In an LBA model of a same/difference task, evidence is collected in 'same' and 'different' accumulators until a threshold level of evidence is collected, which triggers the corresponding response. The time taken for evidence accumulation is the *decision* time, and is combined with a *non-decision* time to give the observed response time. The non-decision time is thought to comprise the time taken to encode the stimuli and make the required motor response.

Does a single process model capture the data?

At the beginning of the paper we described a concern with using response deadline tasks as a tool for investigating the time course of similarity comparisons. As illustrated in Figure 2, it is entirely possible to obtain an interaction between type of feature match (MIPs versus MOPs) and response deadline on overall accuracy – of precisely the kind observed by Goldstone and Medin (1994) – even when the relative importance of MIPs and MOPs does not change during stimulus processing. This is problematic for theoretical models such as SIAM (Goldstone, 1994) that predict that structural alignment processes operate more slowly than simple feature matching processes. One of our primary goals in switching to an instructional manipulation was to conduct a stronger empirical test, one capable of providing clearer evidence that the time course of similarity does indeed involve a structure matching process that operates more slowly than fast feature matching.

To that end, we used the LBA framework to develop several "time-homogeneous" models of similarity. The critical characteristic of these models is the assumption that – regardless of how it is computed – similarity does not change during the time course of the decision. Such a model can allow MIPs to contribute more than MOPs to the subjective similarity, and can allow them to combine in complicated ways, but cannot allow the relative importance of MIPs and MOPs to change within a decision.

The structure of a similarity-driven LBA model is depicted in Figure 5. We assume a single, time-homogeneous similarity signal s that drives two accumulators, one corresponding to the 'same' response and the other corresponding to the 'different' response. The evidence in each accumulator at the start of a trial is sampled from a uniform distribution with range 0 to A. Over time, the evidence increases linearly in both accumulators, where the rate of evidence accumulation v is referred to as the drift rate. In our framework, both accumulators are tied to the similarity between items: the mean drift rate in the respond-same accumulator is equal to the stimulus similarity s, whereas for the respond-different accumulator the mean drift rate is equal to the dissimilarity $1 - s$. The drift rates vary from trial to trial, and are sampled from a normal distribution with standard deviation σ.

The decision making process in an LBA model is simple: as soon as one of the two accumulators reaches a threshold b, that accumulator terminates the decision and an appropriate choice is made. The response time is the sum of the decision time (time taken to reach threshold) and a non-decision time t_{nd}, intended to capture the time taken for the decision process to commence as well as the time taken to produce a motor response. For simplicity, non-decision time is assumed to be a fixed length.

To illustrate how strongly our data constrain the possible models one might propose, we begin by considering a flexible *single source LBA* model, one that allows the stimulus similarity to vary freely across conditions. In the single source LBA model, there need not be any particular relationship between the similarity in any two conditions. Rather, the model is given complete freedom to estimate any pattern of similarities that would let it fit the empirical data. As such there are six similarity parameters, one for each stimulus condition: $s_{0[0]}$, $s_{0[2]}$, $s_{2[0]}$, $s_{0[4]}$, $s_{2[2]}$, $s_{4[0]}$. These six similarities produce the average drift rate v or $1 - v$ for all accumulators, but the trial-to-trial variability σ is a free parameter, as is the variability in start point A. As is traditional in response time modeling, we assume that the instructional manipulation changes people's response threshold parameters, so the model contains a threshold parameter b_a for the *accuracy* condition and another threshold b_s for the *speed* condition. Finally, the single source LBA model allows the non-decision time to differ as a function of instruction condition, yielding two additional parameters $t_{nd,a}$ and $t_{nd,s}$. In total, the single source model uses 12 free parameters to capture the joint distribution over response times and accuracies in all 12 conditions. This and all models were fit independently for each participant using maximum-likelihood parameter estimation after the fastest and slowest 2.5% of responses across all participants were discarded.

How well does the single source LBA capture human performance? It turns out that this model performs poorly, as illustrated in Figure 6. Though we fit all models to the full joint distribution of choices and response times, the failure of the model is apparent in

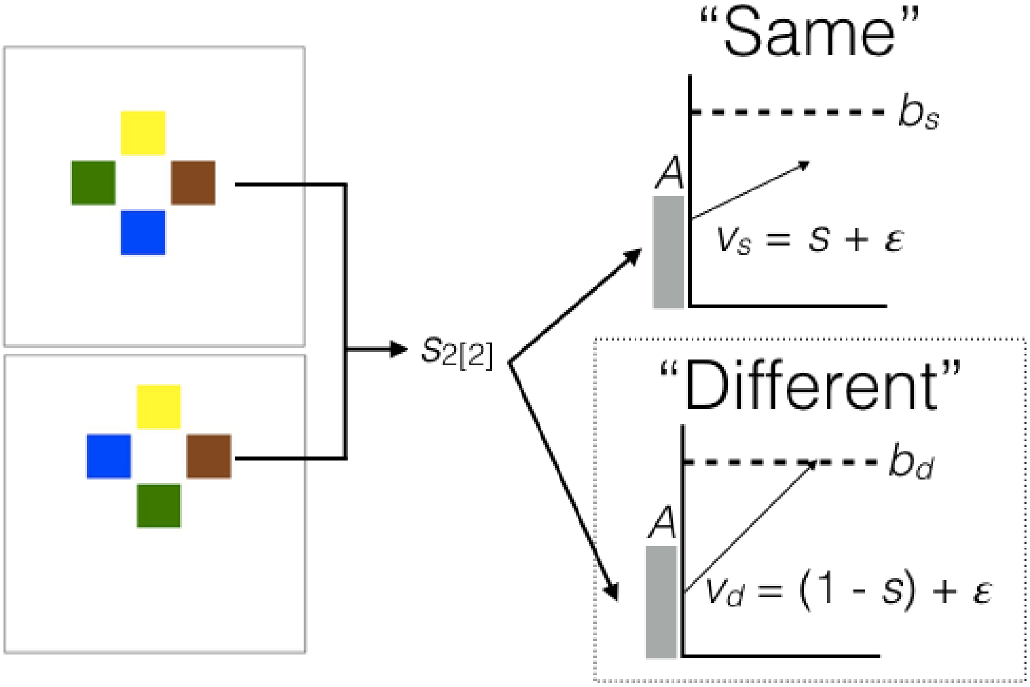

Figure 5. A single process LBA model for similarity-based decisions. The similarity of the stimulus type (s) determines the drift rate for the 'same' response accumulator (top row, v_s) and the 'different' response accumulator (bottom row, v_d). Each accumulator randomly starts at a height between 0 and A and the choice is determined by which accumulator reaches the boundary (b_s or b_d) first ('different' in this example). The reaction time is the duration of accumulation plus the non-decision time parameter ($t_{nd,s}$ or $t_{nd,d}$).

the predicted choice probabilities. In the left panel of Figure 6, the proportion of different responses are plotted as a function of similarity condition. The observed data are plotted as filled and open circles, while the model predictions are plotted as crosses.

Even without a formal statistical evaluation, it is apparent that the single source LBA model shows systematic departures from the empirical data, and is not a plausible account of human behavior in this task. The model fails to capture the difference between speed and accuracy emphasis conditions, especially in the 2[2] stimulus condition, where participants fell to below-chance performance under *speed* instructions. This problem is particularly remarkable given that we allowed similarity to vary freely, and did not constrain it to be any particular function of the number of MIPs and MOPs in the stimulus. The only constraint we imposed is that did we not allow the similarity to change during the time course of processing.

Other single process models

To illustrate why the single-source LBA model behaves so poorly, it is instructive to consider two simpler models, a *featural LBA* in which stimulus similarities depend only on the number of feature matches (i.e., the number of MIPs plus the number of MOPs)

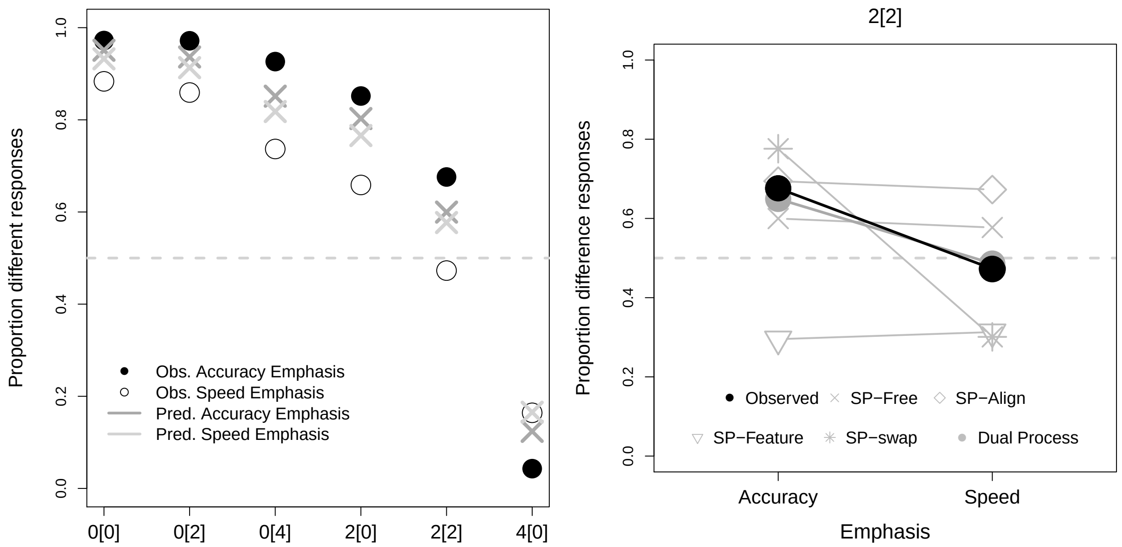

Figure 6. The proportion of different responses observed and predicted by a range of models. Left panel: The predictions of the single source LBA model in which similarity is estimated freely from data (gray crosses) fail to capture the observed data (open and filled circles). Right panel: Focusing on the 2[2] stimulus condition, we see that only the dual process model (filled gray circle) is capable of capturing the below-chance performance in the speed-emphasis condition. All of the other single process models fail to capture this particular pattern.

and a *strict-alignment LBA* model that relies only upon aligned feature matches (i.e., only MIPs) when assessing stimulus similarity. The featural LBA and the strict-alignment LBA both place strong constraints on how similarity is computed. The featural LBA does not distinguish between MIPs and MOPs, and as a consequence is incapable of responding differently in the 0[4], 2[2] and 4[0] conditions: the drift rates for all three of these conditions are constrained by a single featural similarity parameter s_{f4}. By analogy, the s_{f2} parameter defines the stimulus similarity in the 0[2] and 2[0] conditions, and the s_{f0} parameter specifies similarity for the 0[0] condition. The strict-alignment LBA is defined in much the same way, based only on the MIP count: the s_{a0} parameter is used in the 0[0], 0[2] and 0[4] conditions, the s_{a2} parameter is relevant to the 2[0] and 2[2] conditions, and the s_{a4} parameter is used only in the 4[0] condition.

The right panel of Figure 6 highlights some key failures of the single-source models. The right panel plots the proportion of different responses in the speed and accuracy emphasis conditions in the 2[2] stimulus condition for observed data (solid, filled characters) and model predictions (gray characters). The feature match model (gray triangles), performs below chance because it can not distinguish between feature matches that are in or out of place. In contrast, the alignment model is unable to make systematic errors when forced to respond quickly, because it is not ‘tricked’ by the out-of-place feature matches.

Taken together, the failures of these three single process models are instructive. The featural LBA model cannot capture differential sensitivity to MIPs over MOPs, and the strict-alignment LBA ignores the way in which MOPs can be very diagnostic. Though the original single-source LBA model can predict an interaction in a deadline task (Figure 2) it fails to adequately describe human performance across all conditions. This seems to be because feature match information and alignment information seem to play different roles in shaping choices. Specifically, these sources of information do not necessarily combine together into a single, stable similarity score, instead they seem to become available at different time points in the decision process. A more flexible decision-making architecture is required.

Does a dual process model perform better?

The failure of all three single process models to capture the key trends in the data is revealing, and suggests that a theoretical account of similarity-based decision making needs to be able to describe a fundamental shift in perceived stimulus similarity during the time course of the decision. For example, the SIAM model proposed by Goldstone (1994) describes a connectionist system that actively constructs the relational mappings between stimuli as the decision making process unfolds. Because the mappings between the slots in different objects do not initially exist, the model does not at first make a clear distinction between MIPs and MOPs, but as it settles on a preferred way of mapping the two objects, MIPs begin to exert more influence on the similarity between items.

One difficulty in applying SIAM directly is that is not a choice response time model: it does not have a mechanism for terminating the stimulus processing, much less a means to strategically adjust decision criteria when given different instruction sets. However, we can use the qualitative principles in SIAM as a way to motivate a *dual process LBA* model that incorporates a systematic shift from a simple feature matching process to one that weighs matches in place more heavily.

The architecture of the dual process LBA is shown in Figure 7. The model consists of two pairs of 'same' and 'different' accumulators, one for each different kind of similarity. For the feature match accumulators, the similarity signal completely disregards the structural information and treats MIPs and MOPs identically (as per the featural LBA model). In contrast, the structure match accumulators only consider a feature match if there is a relational mapping between the two slots: it only considers MIPs (as per the strict-alignment LBA). The architecture of the model allows *any* of the four accumulators to trigger a response and terminate the decision independently of any other.

The components of the dual process model proceeds much the same as with the previous single process LBA models. The three parameters used to describe the featural similarity accumulators: s_{f0}, s_{f2} and s_{f4}, match the featural LBA. The three parameters to specify similarities based only on the MIP count: s_{a0}, s_{a2} and s_{a4}, match the strict-alignment LBA. The drift rates v for the respond-same and respond-different accumulator are both constructed from the relevant similarities as per Figure 5. We assume that the evidence required to make a decision may be different depending on whether the evidence is based on feature or alignment information (e.g., people might be especially loathe to make feature-match decisions under accuracy instructions given that the correct answer often depends on structural information). Accordingly the model has four separate threshold parameters, two for the accuracy condition – denoted b_{fa} for the feature-match accumulator and b_{aa} for the alignment-match accumulator – and two for the speed condition (b_{fs} and b_{as}). Moreover, we assume that the feature and alignment processes may begin at different times, and do

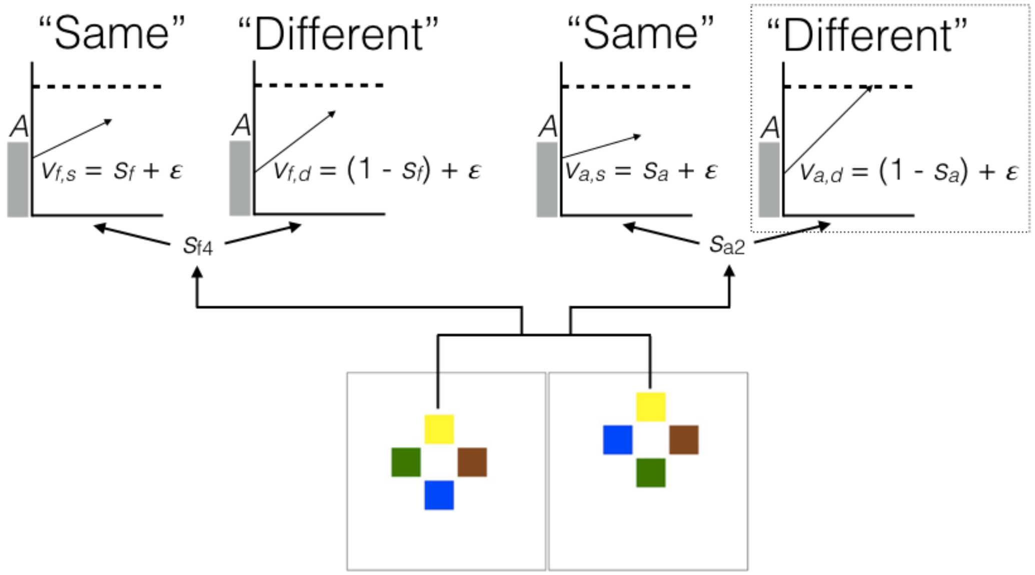
“Same”
“Different”
“Same”
“Different”
A
A
A
A
$v_{f,s} = s_f + \varepsilon$
$v_{f,d} = (1 - s_f) + \varepsilon$
$v_{a,s} = s_a + \varepsilon$
$v_{a,d} = (1 - s_a) + \varepsilon$
s_{f4}
s_{a2}

Figure 7. A dual process LBA model architecture that captures time-inhomogeneous similarity by postulating two separate processes. The drift rates for one set of processes is determined by the featural similarity (MIPs + MOPs) of the stimuli (left set) while the drift rate for the right set is determined only by aligned features (MIPs).

so by allowing for a separate non-decision time for each source of evidence ($t_{nd,a}$ and $t_{nd,f}$). As before, we have a single parameter A that governs start point variability and another parameter σ that governs drift rate variability. In total the dual process LBA model uses 14 free parameters to describe 12 joint probability distributions over response time and accuracy.

How well does this dual process LBA model perform? The right panel of Figure 6 shows that, unlike the single process models described previously, the dual-process model (filled gray circles) can produce the dramatic change in performance across emphasis conditions in the `2[2]` stimulus condition. It captures the systematic mis-identification of `2[2]` stimuli as being the same while under speed emphasis (like humans it performs slightly below chance on these items), but is able to reverse the predicted pattern under accuracy emphasis.

The dual-process model also provides an impressive account of the full distribution of response times and choices in all 12 conditions, as shown in Figure 8. Each panel in the figure is a plot of the cumulative density function (CDF) for the probability of responding across five quantiles of the full distribution (.1, .3, .5, .7, and .9). Lines connect the points from the same response distribution where the x-coordinate indicates the cumulative probability of a given response and the y-coordinate indicates the median response time in that quantile. The empirical data are plotted in black, with filled points for correct responses and open points for incorrect responses. To interpret the CDF plots, note that higher accuracy is reflected in the height of the filled circles, and the speed of responses are indicated by the steepness of the curve. Model predictions are plotted in gray. Given the difficulty that all

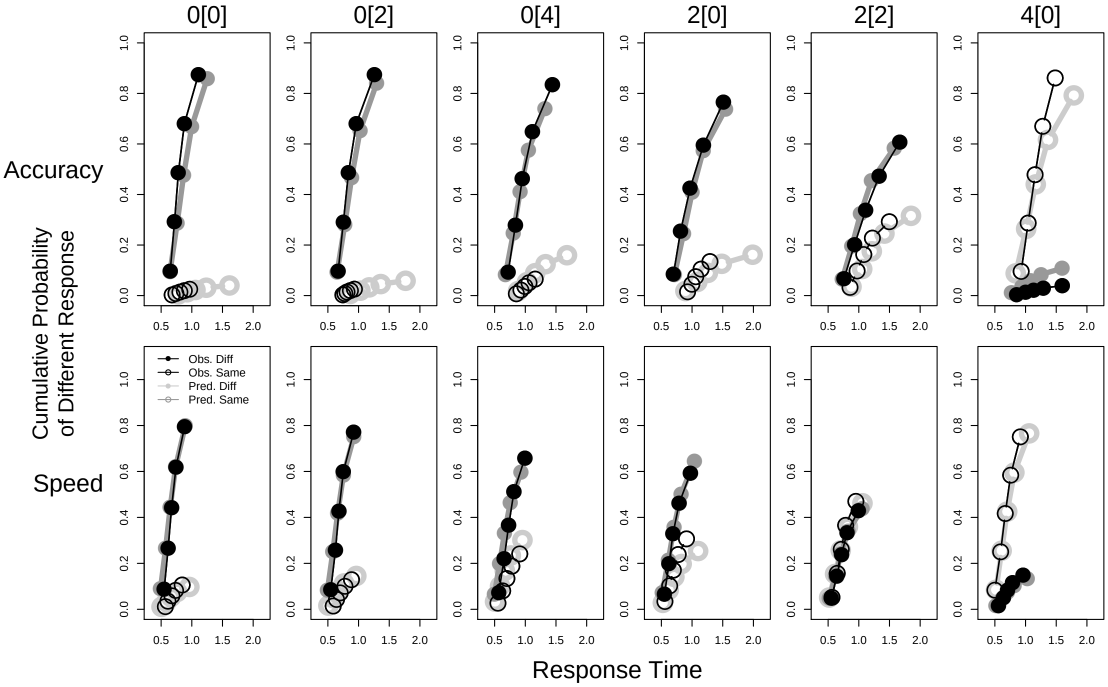

Figure 8. Performance of the dual-process LBA model. This model is composed of four accumulators for each decision: separate feature-based and alignment-based accumulators for both same (filled points) and different (open points) judgments. The black lines indicate human performance and the gray lines indicate model predictions.

single-process models had with accounting for the data, we take the good performance of this dual-process model as rather convincing evidence of the important role of both sources of evidence.

Can source switching explain the results?

The impressive performance of the dual process LBA model raises the question of whether it would be possible to produce human-like behavior in a simpler fashion. The key claim in the dual process LBA is that simple feature matching information and more complicated alignment-based knowledge interact with one another during the time course of a single decision. However, given the concerns we raised initially with the response deadline task in which a time-homogeneous model was able to produce human like behavior, it is important to consider other possible explanations for the data. We have already considered three different single source models, all of which fail to reproduce the qualitative characteristics of the data. We now consider a fourth possibility.

One possible account of the shift from simple feature matching to alignment based decisions might be to suggest that the shift only occurs between instructional conditions. Participants might rely solely on feature match information in the *speed* condition, and switch to aligned feature matches in the *accuracy* condition. The corresponding *source-switching LBA* model is a "dual process" model in name only: while it does allow for people to be sensitive to both MIPs and MOPs in different ways, it does not allow any changes to occur during a decision.

Can such a model account for our data? As shown in the right panel of Figure 6, this model performs performs poorly and overestimates the difference between the Accuracy and Speed conditions for the `2[2]` stimulus type.

What does the dual-process LBA model tell us?

The parameters in the dual-process model, plotted in Figure 9, appear to do sensible things. The average evidence accumulation rates for the 'different' accumulator decrease as the number of feature-matches increases between the two items (top left) The evidence for a difference also decreases as there are more matches in place (top right). Also as expected, the non-decision time is faster for the feature-based accumulators than the alignment-based accumulators (bottom right). Finally, the threshold is higher and thus more conservative in accuracy instruction conditions than speed instruction conditions (bottom left). Taken together, the parameter estimates from the dual-process LBA model strengthen the claim that the nature of the similarity signal changes during the time course of comparison, with feature match information arriving early and relational information appearing later.

Beyond providing support for the claim that perceived similarity changes during a decision process, the dual process LBA model allows us to examine how people strategically adjust their decision policy in response to instructional manipulations. As the bottom left panel of Figure 9 shows, when asked to emphasize speed people adjust the decision threshold for *both* accumulators to be very low, producing fast decisions but introducing more errors. Moreover, because both thresholds are set at similar levels, the model is able to trigger decisions using either raw feature match information or by using structural information based on MIPs only. In contrast, when the instructions emphasize accuracy, both thresholds increase, but the rise is asymmetric: the response threshold for the feature match information is set much higher than for the alignment information. The consequence of this is that under accuracy instructions the decision is much more likely to be triggered by the aligned feature accumulator than the raw feature accumulator. In effect, under accuracy instructions people systematically adjust their decision process such that more decisions are made based on the number of aligned features.

General Discussion

Over the last decade there has been a steady stream of theoretical and empirical results suggesting that the evidence for "dual process" accounts of cognitive phenomena is weaker than it appears at face value (Newell & Dunn, 2008; Dunn, 2008, 2004; Lassiter & Goodman, 2015; Stephens et al., in press; Newell et al., 2011). Very typically, a phenomenon that appears to reflect the operation of multiple qualitatively distinct process turns out on closer inspection to reflect simple criterion shifts in response to difficulty or instructional demands (Dunn, 2008; Dube et al., 2010). The implication for theories of stimulus similarity would be substantial, were the same pattern to appear in this domain. Fortunately, in this context at least, this does not appear to be the case: even with relatively simple stimuli and simple judgments, it is difficult to see how to account for the empirical data without a fast-operating feature matching process and a slower structure alignment process of some kind.

Besides providing a stronger assurance that there is a genuine shift from featural match to aligned feature matches during comparison, our findings highlight the role that strategic decision processes play in simple stimulus comparisons. When the task demands shift from an speed emphasis to an accuracy emphasis, people do not merely exercise more caution when making decisions, they shift the *emphasis* from feature matches to structural information, as highlighted by the asymmetric change in decision thresholds.

More broadly, the success of the dual process LBA model opens up a number of questions. Firstly, how can we integrate the decision architecture used in the LBA model with a more theoretically oriented similarity model such as SIAM? The strength of SIAM and related models (e.g., Larkey & Love, 2003) is that they explicitly model the process by which the learner *constructs* the mapping between different stimuli, and in doing so they provide an account of how the learner comes to differentiate between MIPs and MOPs, and even to have a more nuanced perception of what counts as a good match than a simple MIP and MOP count provides. Our LBA model is silent on how these processes operate, and merely stipulates that there must exist *some* alignment process that differentially processes MIPs and MOPs and best-fitting non-decision time parameters indicate this process comes online later than feature information in the decision process. Clearly, as a theoretical account of how similarity is constructed, the LBA model is somewhat deficient. However, what it gains in exchange for this simplicity is the ability to describe a more complicated decision making architecture. In our LBA model, the learner is given the freedom to decide for themselves *when* they have enough evidence to justify making a response, and as it turns out people are very sophisticated in the way they adjust the decision policy in response to different task demands. Indeed, while – in retrospect – the pattern of decision thresholds shown in Figure 9 is intuitive, it was not (at least to us) obvious a priori that the striking pattern of choices and response times (Figure 4) could be captured (Figure 8) in such an elegant way. The systematic shift in RT across stimulus conditions under accuracy instructions coupled with perfectly flat pattern of median RT under speed instructions is somewhat unusual. When these RT patterns are paired with the qualitative shift in the `2[2]` condition, the data set as a whole provides a very strong constraint on possible models, one that appears to be impossible to capture with a single process model in the LBA framework, yet perfectly natural to accommodate with a dual process model. A natural direction for future research would be to develop models that combine the strengths of both approaches, with a detailed (SIAM-like) mapping process that feeds into a flexible (LBA-like) decision architecture.

Another open question pertains to the uniqueness of the decision architecture. From a similarity modeling perspective, the key theoretical question is when structural information becomes available to guide similarity-based decisions, and our dual process LBA model provides one useful way to describe the phenomenon. However it does so using a parallel architecture that treats feature match information and match-in-place information as separate processes, each with its own decision threshold. Using the language of response time models, it is a parallel self-terminating decision model that uses multiple time-homogeneous accumulators, but this is not the only way that the two sources of evidence could be combined. For instance, if we were to “read off” the similarity produced by the SIAM model during the decision process, we would end up with a rather different architecture, one in which the two sources of evidence are "merged" into a single time-*inhomogenous* signal. It is not clear which of these two approaches provides the better account of human similarity comparisons, nor is it obvious whether these different architectures are empirically discriminable (see, e.g., the serial-parallel mimicry problem, Algom, Eidels, Hawkins, Jefferson, & Townsend, 2015). The dual process LBA account that we introduce provides one way of accounting for similarity-based decisions, and our experimental data rule out a variety of alternative single process accounts, but other possibilities exist.

As a final note, the success of the LBA based framework opens up a new possible line of work. As a methodological tool, the dual process architecture allows us to examine when different sources of information become available to the decision maker. In these experiments we have focused on simple feature match information versus structural match in place information, but the approach could potentially be applied in other contexts. For instance, in a preliminary investigation Hendrickson, Navarro, and Donkin (2015) found evidence suggesting that the conflict between thematic and taxonomic information (Lin & Murphy, 2001) that unfolds over development (Piaget & Inhelder, 1964; Markman, 1989) appears to be mirrored during the time course of a single decision (see also, Gentner & Brem, 1999), but further investigation would be needed to justify a strong claim to that effect. More generally, we argue that there is considerable benefit to be obtained from modeling the decision processes involved when people compare the similarity between different items, and to that end the field might benefit from using state of the art decision models such as the LBA.

Acknowledgments

DJN received salary supported from ARC grant FT110100431. Research costs and salary support for ATH were funded through ARC grant DP110104949. Portions of this research were presented at the *37th Annual Conference of the Cognitive Science Society.*

References

Algom, D., Eidels, A., Hawkins, R. X., Jefferson, B., & Townsend, J. T. (2015). Features of response times: Identification of cognitive mechanisms through mathematical modeling. *Oxford library of psychology. The Oxford handbook of computational and mathematical psychology*, 63–98.

Brown, S. D., & Heathcote, A. (2008). The simplest complete model of choice response time: Linear ballistic accumulation. *Cognitive Psychology*, *57*(3), 153–178.

Dube, C., Rotello, C. M., & Heit, E. (2010). Assessing the belief bias effect with ROCs: It's a response bias effect. *Psychological Review*, *117*(3), 831–863.

Dunn, J. C. (2004). Remember-know: a matter of confidence. *Psychological Review*, *111*(2), 524–542.

Dunn, J. C. (2008). The dimensionality of the remember-know task: a state-trace analysis. *Psychological Review*, *115*(2), 426–446.

Falkenhainer, B., Forbus, K. D., & Gentner, D. (1989). The structure-mapping engine: Algorithm and examples. *Artificial Intelligence*, *41*(1), 1–63.

Gentner, D. (1983). Structure-mapping: A theoretical framework for analogy. *Cognitive Science*, *7*(2), 155–170.

Gentner, D. (1989). The mechanisms of analogical learning. *Similarity and Analogical Reasoning*, *199*, 199–241.

Gentner, D., & Brem, S. (1999). Is snow really like a shovel? distinguishing similarity from thematic relatedness. In *Proceedings of the twenty-first annual meeting of the Cognitive Science Society* (p. 179-184).

Goldstone, R. L. (1994). Similarity, interactive activation, and mapping. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *20*(1), 3–28.

Goldstone, R. L., & Medin, D. L. (1994). Time course of comparison. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *20*(1), 29–50.

Goldstone, R. L., Medin, D. L., & Halberstadt, J. (1997). Similarity in context. *Memory & Cognition*, *25*(2), 237–255.

Goodman, N. (1972). Seven strictures on similarity. In *Problems and projects.* Bobs-Merril.

Hawkins, G. E., Hayes, B. K., & Heit, E. (2016). A dynamic model of reasoning and memory. *Journal of Experimental Psychology: General*, *145*(2), 155–180.

Hendrickson, A., Navarro, D. J., & Donkin, C. (2015). Quantifying the time course of similarity. In *Proceedings of the 37th annual meeting of the Cognitive Science Society* (pp. 908–913).

Hummel, J. E., & Holyoak, K. J. (1997). Distributed representations of structure: A theory of analogical access and mapping. *Psychological Review*, *104*(3), 427–466.

Larkey, L. B., & Love, B. C. (2003). Cab: Connectionist analogy builder. *Cognitive Science*, *27*(5), 781–794.

Lassiter, D., & Goodman, N. D. (2015). How many kinds of reasoning? inference, probability, and natural language semantics. *Cognition*, *136*, 123–134.

Lin, E. L., & Murphy, G. L. (2001). Thematic relations in adults' concepts. *Journal of Experimental Psychology: General*, *130*(1), 3–28.

Luce, R. D. (1986). *Response times.* Oxford University Press.

Markman, E. M. (1989). *Categorization and naming in children: Problems of induction.* Mit Press.

Medin, D. L., Goldstone, R. L., & Gentner, D. (1990). Similarity involving attributes and relations: Judgments of similarity and difference are not inverses. *Psychological Science*, *1*(1), 64–69.

Medin, D. L., Goldstone, R. L., & Gentner, D. (1993). Respects for similarity. *Psychological Review*, *100*(2), 254–278.

Morey, R. D., & Rouder, J. N. (2015). Bayesfactor: Computation of bayes factors for common designs [Computer software manual]. Retrieved from `https://CRAN.R-project.org/package=BayesFactor` (R package version 0.9.12-2)

Newell, B. R., & Dunn, J. C. (2008). Dimensions in data: Testing psychological models using state-trace analysis. *Trends in Cognitive Sciences*, *12*(8), 285–290.

Newell, B. R., Dunn, J. C., & Kalish, M. (2011). Systems of category learning: Fact or fantasy? *Psychology of Learning and Motivation-Advances in Research and Theory*, *54*, 167–215.

Nosofsky, R. M. (1986). Attention, similarity, and the identification–categorization relationship. *Journal of Experimental Psychology: General*, *115*(1), 39–57.

Piaget, J., & Inhelder, B. (1964). *The early growth of logic in the child: Classification and seriation (ea lunze & d. papert, trans.).* London: Routledge & Kegan Paul.

Ratcliff, R. (1978). A theory of memory retrieval. *Psychological Review*, *85*(2), 59–108.

Ratcliff, R. (2006). Modeling response signal and response time data. *Cognitive Psychology*, *53*(3), 195–237.

Ratcliff, R., & Smith, P. L. (2004). A comparison of sequential sampling models for two-choice reaction time. *Psychological Review*, *111*(2), 333–367.

Rouder, J. N., Morey, R. D., Speckman, P. L., & Province, J. M. (2012). Default bayes factors for anova designs. *Journal of Mathematical Psychology*, *56*(5), 356–374.

Smith, L. B., & Heise, D. (1992). Perceptual similarity and conceptual structure. *Advances in Psychology*, *93*, 233–272.

Stephens, R. G., Dunn, J. C., & Hayes, B. K. (in press). Are there two processes in reasoning? the dimensionality of inductive and deductive inferences. *Psychological Review.*

Treisman, A., & Gelade, G. (1980). A feature-integration theory of attention. *Cognitive Psychology*, *12*(1), 97–136.

Treisman, A., & Schmidt, H. (1982). Illusory conjunctions in the perception of objects. *Cognitive Psychology*, *14*(1), 107–141.

Tversky, A. (1977). Features of similarity. *Psychological Review*, *84*(4), 327–352.

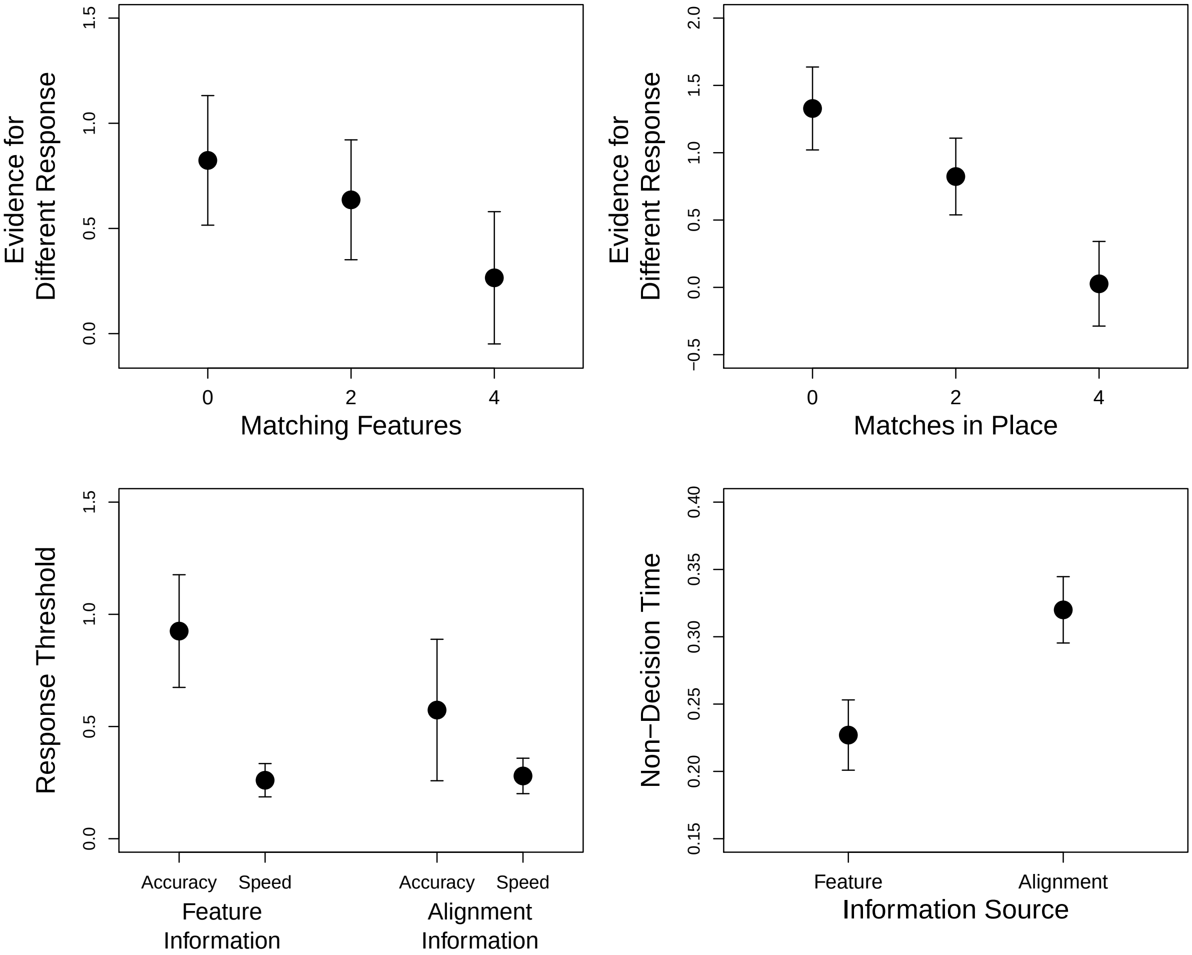

Figure 9. Parameter estimates for the dual process LBA model. The top row shows the accumulation rates for different responses for feature matches (left) and matches in place (right) decrease as there are more matches. The bottom left panel shows that the decision boundaries are placed higher under accuracy conditions than speed conditions, for both feature and alignment information. Finally, the bottom right panel suggests that the onset time for the alignment information is later than for feature information. Error bars indicate 1 standard error over participants.