Cognitive Psychology 71 (2014) 27–54

Contents lists available at ScienceDirect

Cognitive Psychology

journal homepage: www.elsevier.com/locate/cogpsych

The discovery and comparison of symbolic magnitudes

Dawn Chen [a,*], Hongjing Lu [a,b], Keith J. Holyoak [a]

[a] Department of Psychology, University of California, Los Angeles, 1285 Franz Hall, Box 951563, Los Angeles, CA 90095-1563, United States
[b] Department of Statistics, University of California, Los Angeles, United States

ARTICLE INFO

Article history:
Accepted 10 January 2014

Keywords:
Magnitude comparisons
Reference points
Semantic congruity
Symbolic distance
Markedness

ABSTRACT

Humans and other primates are able to make relative magnitude comparisons, both with perceptual stimuli and with symbolic inputs that convey magnitude information. Although numerous models of magnitude comparison have been proposed, the basic question of how symbolic magnitudes (e.g., size or intelligence of animals) are derived and represented in memory has received little attention. We argue that symbolic magnitudes often will not correspond directly to elementary features of individual concepts. Rather, magnitudes may be formed in working memory based on computations over more basic features stored in long-term memory. We present a model of how magnitudes can be acquired and compared based on BARTlet, a representationally simpler version of *Bayesian Analogy with Relational Transformations* (BART; Lu, Chen, & Holyoak, 2012). BARTlet operates on distributions of magnitude variables created by applying dimension-specific weights (learned with the aid of empirical priors derived from pre-categorical comparisons) to more primitive features of objects. The resulting magnitude distributions, formed and maintained in working memory, are sensitive to contextual influences such as the range of stimuli and polarity of the question. By incorporating psychological reference points that control the precision of magnitudes in working memory and applying the tools of signal detection theory, BARTlet is able to account for a wide range of empirical phenomena involving magnitude comparisons, including the symbolic distance effect and the semantic congruity effect. We discuss the role of reference points in cognitive and social decision-making, and implications for the evolution of relational representations.

* Corresponding author. Fax: +1 310 206 5895.
E-mail addresses: sdchen@ucla.edu (D. Chen), hongjing@ucla.edu (H. Lu), holyoak@lifesci.ucla.edu (K.J. Holyoak).

0010-0285/$ - see front matter © 2014 Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.cogpsych.2014.01.002

1. Introduction

Humans and other primates have sophisticated abilities to learn and make judgments based on relative magnitude. Magnitude comparisons are critical in making choices (e.g., which of two products is more desirable?), making social evaluations (e.g., which person is friendlier?), and in many other forms of appraisal (e.g., who can run faster, this bear or me?). In addition to making comparisons based on elementary perceptual dimensions (e.g., identifying the longer of two line segments or the brighter of two lights), people are able to make analogous judgments based on symbolic dimensions using information stored in memory (e.g., the relative size or intelligence of various animals). Non-human primates are also capable of at least rudimentary symbolic comparisons. For example, rhesus monkeys are capable of learning shapes (Arabic numerals) that correspond to small numerosities (1–4 dots), such that the shapes acquire neural representations overlapping those of the corresponding perceptual numerosities and can be compared on that basis (Diester & Nieder, 2007).

Striking parallels have been observed between perceptual and symbolic judgments. In particular, both perceptual and symbolic judgments yield a *distance* effect, such that the ease of judgments (indexed by accuracy and/or reaction time) increases with the magnitude difference between the objects being compared (e.g., Moyer, 1973; Moyer & Bayer, 1976; Moyer & Landauer, 1967). A symbolic distance effect is observed not only with quasi-perceptual dimensions such as size, but also with more abstract dimensions such as animal intelligence (Banks, White, Sturgill, & Mermelstein, 1983) and with scalar adjectives of quality (e.g., *good*, *fair*; Holyoak & Walker, 1976). Non-human primates also exhibit a distance effect for judgments along various perceptual dimensions, including numerosity (Nieder & Miller, 2003).

When judgments are made using contrastive polar concepts (e.g., "choose brighter" versus "choose dimmer", "choose better" versus "choose worse"), both perceptual (Audley & Wallis, 1964; Petrusic & Baranski, 1989; Wallis & Audley, 1964) and symbolic judgments also yield a *semantic congruity* effect: for objects with high values on the dimension, it is easier to judge which object is greater, whereas for objects with low values, it is relatively easier to judge which is lesser (e.g., Banks, Clark, & Lucy, 1975; see Moyer & Dumais, 1978, for an early review). Like the distance effect, semantic congruity effects have also been obtained with monkeys (Cantlon & Brannon, 2005). A further phenomenon, the *markedness* effect, refers to the fact that for some pairs of polar adjectives, one (the "unmarked" form) is easier to process overall than the other (Clark, 1969). For example, the "unmarked" question "Which is larger?" tends to be answered more rapidly overall than the "marked" question "Which is smaller?" (Clark, 1969; Clark, Carpenter, & Just, 1973). The impact of markedness implies that the congruity effect often takes the form of an asymmetrical interaction.

1.1. How are magnitudes generated?

Numerous models of symbolic magnitude comparisons have been proposed, and we will review several of them below. However, in the present paper we focus on a question that (even though it is arguably the most basic of all) has seldom been asked, far less answered: where do subjective magnitudes come from? In the case of perceptual judgments with unidimensional stimuli (e.g., tones varying in loudness), it is reasonable to assume that a specific neural channel generates magnitudes. For symbolic comparisons, the tacit assumption has been that the long-term memory representation of each object being compared includes a magnitude value (perhaps with an associated variance), and that these magnitudes are simply retrieved and loaded into working memory, where a comparison process operates.

For a few types of symbolic comparisons, such as numerical magnitudes of digits, it may indeed be the case that each object has a pre-stored magnitude in long-term memory. But for more complex dimensions this assumption is questionable, and indeed quite unrealistic. Even symbolic size judgments, which are closely linked to perceptual features, are unlikely to always be based on pre-stored magnitudes, as size is actually a complex function of three-dimensional shape. Indeed, recent evidence indicates that although numerical magnitudes are automatically activated when reading integers, size

magnitudes associated with animal names are activated only when the reader has the goal of making size comparisons (Hoedemaker & Gordon, 2013). People may have stored size values for a few "landmark" objects (e.g., an elephant or a mouse), but are unlikely to have pre-stored size values for less familiar animals (e.g., a beaver or a swordfish). The notion that magnitudes are pre-stored becomes yet more implausible for the wide range of dimensions on which people can make symbolic comparisons, especially in the interpersonal and social realm (e.g., intelligence, friendliness, religiosity, conservatism). Rather than being elementary components of concept meanings, magnitudes may often be derived, context-dependent features (Goldstone, 1994; Smith, Gasser, & Sandhofer, 1997). Furthermore, rather than being pre-stored, magnitudes may be computed as needed in response to a query.

It follows that a comprehensive account of symbolic magnitude comparisons must begin with a model of how symbolic magnitudes are discovered. One general hypothesis is that magnitudes can be generated by operations performed on vectors of more elementary features associated with individual objects. Fig. 1 provides a visualization of the sort of input that might underlie people's everyday knowledge of various types of animals. These representations were derived from norms of the frequencies with which participants at the University of Leuven generated features characterizing various animals (De Deyne et al., 2008; see Shafto, Kemp, Mansinghka, & Tenenbaum, 2011). Each animal in the norms is associated with a set of frequencies across more than 750 features. Fig. 1 includes feature vectors for 30 example animals based on the 50 features most highly associated with a larger set of animal names (Lu et al., 2012). Although these "Leuven vectors" presumably only approximate people's knowledge about animal concepts, they have the great virtue of being derived from independent sources of data, rather than being hand-coded. The simulations reported in the present paper are based on inputs extracted from the Leuven vectors, as well as similar feature vectors created using the topic model (Griffiths, Steyvers, & Tenenbaum, 2007).

Could individual Leuven features be directly used as measures of magnitude? One might have supposed, for example, that the value of the feature "is big" would be sufficient to predict relative size. But although this dimension is indeed the single most important factor predicting size, it is far from sufficient. The Leuven features were derived from the frequencies with which participants generated attributes, and animals for which their large size is salient (often in reference to a subcategory) tended to have higher feature values for "is big" (e.g., based on a comparison of feature values for that attribute alone, the Leuven dataset indicates that an eagle is larger than a giraffe). To address this problem we need distributed representations that can be used to compute derived magnitude dimensions.

To provide such distributed representations, Lu et al. (2012) developed *Bayesian Analogy with Relational Transformations* (BART), a model of how one-place scalar adjectives (e.g., *large*, *smart*) and two-place comparative relations (e.g., *larger*, *smarter*) can be learned from non-relational feature vectors. Using various inputs, including Leuven vectors and vectors derived using the topic model (Griffiths et al., 2007), the model was applied to the acquisition of concepts related to four continuous dimensions: size, ferocity, speed and intelligence. BART incorporates information from a prior probability distribution over a space of weights, as well as examples of animal pairs that instantiate a relation, to obtain a posterior distribution over the weight space, which is used to predict whether the relation holds for novel pairs. Learning is supervised, as the model received training examples that are associated with truth values for the instantiated relation (e.g., the model is told that "cow is larger than dog" is true). Only positive examples of relations are used (since children's concept learning seems to be largely guided by positive examples; see Bloom, 2000). The representations of relational concepts created by BART for each of the four magnitude dimensions of interest turned out to be highly distributed, based on at least 20 statistically predictive features (see Lu et al., 2012, figure 10, pp. 634–635).

The simulation results reported by Lu et al. (2012) suggest that concepts related to symbolic magnitudes can be discovered by inductive learning, rather than simply assumed to be directly available in long-term memory. Moreover, the Bayesian approach in general (and the BART model in particular) implies that magnitudes will be represented not as deterministic values, but rather as probability distributions. The probabilistic framework is in agreement with the intuition that symbolic magnitudes (e.g., the size of a kangaroo, the intelligence of a goat) are "fuzzy" rather than firm, and thus judgments related to these attributes are susceptible to the influence of context.

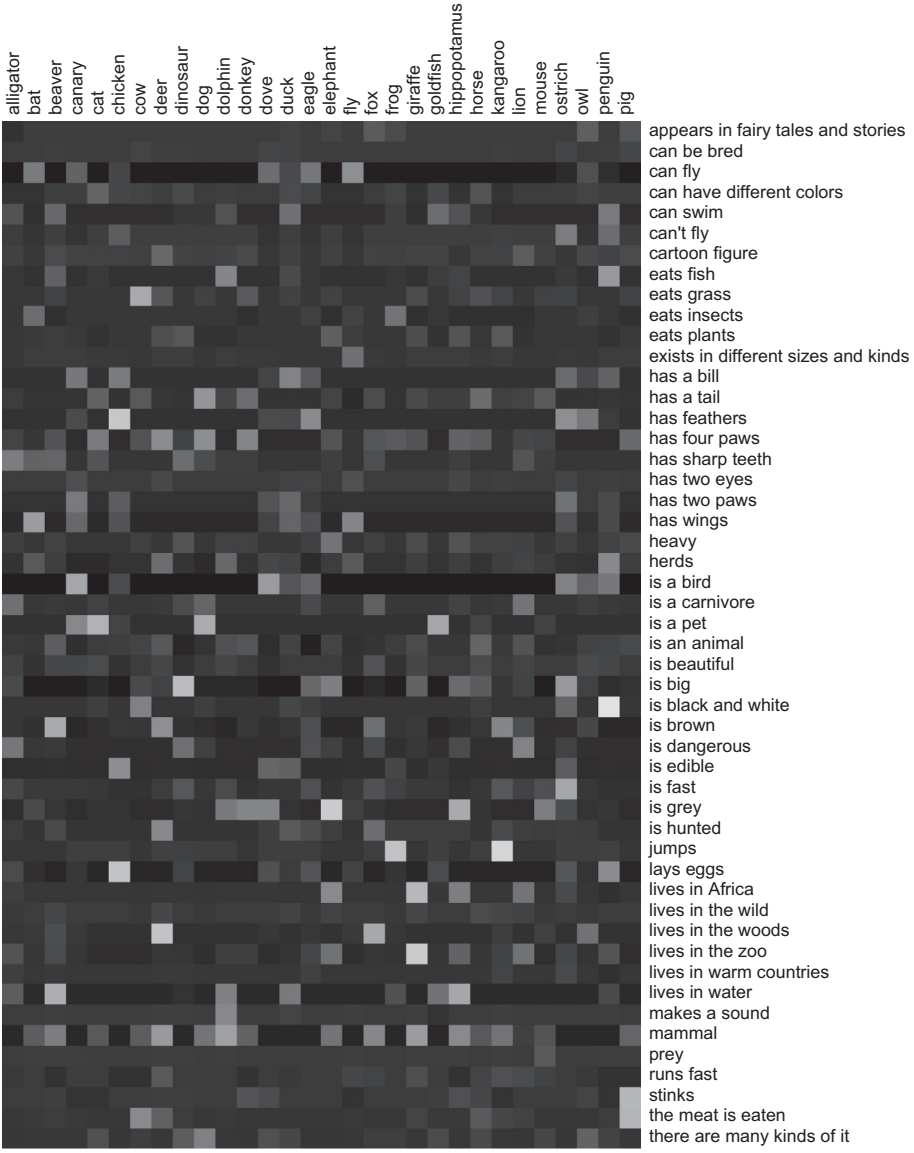

Fig. 1. Illustration of Leuven vectors for some example animals. The cell intensities represent feature values (light indicates high frequency values, dark indicates low frequency values).

Our goal in the present paper is to provide an integrated account of how symbolic magnitudes, represented in working memory as probability distributions, can be created and then used to make comparative judgments. We will first briefly review previous accounts of the three major phenomena observed in studies of comparative judgment: symbolic distance effects, semantic congruity effects, and markedness. We will then show how a model incorporating assumptions about the attentional control of magnitude representations in working memory can provide a unified account of these core phenomena.

1.2. Alternative models of symbolic magnitude comparisons

We will not attempt an exhaustive review of the large literature on mental magnitude comparisons, but rather will focus on findings that give rise to some of the principles we include in our current model (for broader reviews of work with humans see Moyer & Dumais, 1978; Petrusic, 1992; for a review of work with non-human primates see Cantlon, Platt, & Brannon, 2009).

There is virtually complete consensus among current researchers that the ubiquitous distance effect reflects some form of internalized representation of magnitude akin to positions on a number line, such that larger magnitudes are more readily discriminable. This notion goes back at least to Moyer (1973), who referred to an "internal psychophysics" for symbolic comparisons. Behavioral studies have identified striking parallels between symbolic distance effects and those observed in overt perceptual comparisons (e.g., Audley & Wallis, 1964; Holyoak & Patterson, 1981; Moyer & Bayer, 1976). As in the case of perceptual comparisons, the pattern of difficulty for symbolic comparisons suggests that internal magnitudes are typically compressed such that subjective magnitude differences decrease as the absolute magnitudes of the objects being compared increase (Shepard, Kilpatric, & Cunningham, 1975). More recent work has provided strong evidence that humans and other primates are equipped with specialized neural circuitry for dealing with approximate magnitude on various dimensions (e.g., Cantlon, Brannon, Carter, & Pelphrey, 2006; Dehaene & Changeux, 1993; Piazza, Izard, Pinel, Bihan, & Dehaene, 2004; Piazza, Mechelli, Price, & Butterworth, 2006; Piazza, Pinel, Le Bihan, & Dehaene, 2007; Pinel, Piazza, Bihan, & Dehaene, 2004).

Several models for magnitude comparisons have been proposed (for a review see Petrusic, 1992). The evidence distinguishing among them mainly involves the congruity and markedness effects. The congruity effect has been interpreted in multiple ways. An expectancy model (Banks & Flora, 1977; Marschark & Paivio, 1979) assumes that the congruity effect arises because the comparative is presented prior to the stimulus pair, enabling the person to prepare in some way for stimuli within a certain range (e.g., either small or large objects). However, robust congruity effects are found even when the comparative is presented *after* the stimuli to be compared, in a design in which questions about multiple dimensions were intermixed (Holyoak & Mah, 1981). Other studies yielded similar disconfirmatory findings (Banks et al., 1983; Howard, 1983; Shoben, Sailor, & Wang, 1989).

A related explanation of the congruity effect attributes the phenomenon to differential frequency of association between each comparative and items of various magnitudes (i.e., the "greater" comparative may be more often used with items of high magnitude, and the reverse for the "lesser" comparative). However, Rylass and Smith (2000) taught adults novel comparatives, and found that a congruity effect arose even when the training set was designed to eliminate any correlation between the form of the comparative and the magnitude of items. These and other findings concerning acquisition of comparative terms (Ryalls, Winslow, & Smith, 1998) suggest that the congruity effect reflects the meaning of the contrastive terms, rather than unbalanced presentation frequencies during learning that might influence expectancies about items.

A frequency-based explanation has also been offered for markedness effects, as unmarked forms of adjectives are typically used more frequently than the corresponding marked forms. Often the marked term is aptly applied only to the range of magnitudes extending from the negative pole to the midpoint, whereas the unmarked term can be aptly applied across the full magnitude range (Clark, 1969). However, the finding of a markedness effect in monkeys, in a design in which the two forms of the implicit query occurred on an equal number of trials during training, suggests that markedness effects cannot be fully explained by unequal frequency of linguistic use (Cantlon & Brannon, 2005).

The semantic coding model (Banks, Fujii, & Kayra-Stuart, 1976; Banks et al., 1975) attributes the congruity effect to categorical codes based on language (e.g., "large" and "small"). In this model, the congruity effect reflects systematic differences in the probability that the codes for the objects will match the linguistic form of the comparative. Although the model provides a good quantitative fit to some data sets (Banks et al., 1976), it faces a number of problems as a general explanation of symbolic comparisons. Because it is based on linguistic codes, the model is severely strained by the fact that distance, congruity and markedness effects are also observed with non-linguistic primates, such as monkeys (Cantlon & Brannon, 2005; Cantlon et al., 2009). Also, the model cannot explain evidence that similar effects are observed in direct judgments of discriminability among ordered items

(e.g., the form of comparative used in the question influences the relative spacing of cities along an east–west dimension as recovered by scaling methods; Holyoak & Mah, 1982). Finally, the model predicts that the magnitude of the congruity effect will be independent of factors that influence decision difficulty (Banks et al., 1975). However, there is considerable evidence that the magnitude of the congruity effect in fact varies systematically with decision difficulty (Petrusic, 1992; Petrusic & Baranski, 1989; Shaki, Leth-Steensen, & Petrusic, 2006).

1.3. Reference-point models

The final major class of models (and the one most relevant to the present proposal) includes those that locate the congruity and markedness effects within the process of magnitude comparison itself. The intuitive idea is that when judging (for example) whether an elephant is larger than a hippo, the subjective magnitude difference is in fact more discriminable than when judging whether an elephant is *smaller* than a hippo. Such discriminability effects might arise by a mechanism through which the form of the question modulates magnitude representations in working memory. A number of specific models have been proposed, which share the hypothesis that the polarity of the comparative serves to establish a *reference point* at or near the corresponding end of the continuum, and that magnitude differences between objects close to the reference point are discriminated more easily than otherwise comparable differences between objects far from the reference point (Holyoak, 1978; Holyoak & Mah, 1982; Jamieson & Petrusic, 1975; Marks, 1972). Holyoak (1978) argued that attending to a reference point at the congruent extreme of a dimension aids in coding the polarity of the question (i.e., distinguishing between "choose greater" versus "choose lesser" for a specific pair of comparatives).

Reference-point models are not inherently linguistic, and hence can in principle be applied to comparative judgments in non-linguistic species (Cantlon et al., 2009); they can accommodate the influence of the question form on direct discriminability judgments (Holyoak & Mah, 1982); and in some variants (Marks, 1972) they predict the general finding that congruity effects are larger when decisions are more difficult (Petrusic, 1992; see Banks et al., 1975, for a derivation). In addition, reference-point models can potentially explain another critical property of the congruity effect, which is that it is sensitive to the range of magnitudes exhibited in the stimulus set. For example, if the presented stimuli are all relatively small animals (e.g., smaller than a dog), then the *relatively* large animals within this restricted set (e.g., rabbit and beaver) will show an advantage for "choose larger" over "choose smaller" (Čech & Shoben, 1985; Čech, Shoben, & Love, 1990; see also Petrusic & Baranski, 1989). Similar range effects have been observed in studies of comparative judgments by monkeys (Jones, Cantlon, Merritt, & Brannon, 2010). It is natural to suppose that an observer could strategically shift reference points to reflect the magnitude range of the presented stimuli.

A number of explanations of how a reference point exerts its effect have been proposed. Jamieson and Petrusic (1975) and Holyoak (1978) suggested that observers assess the *ratio* of distances from each object to the reference point, rather than simply taking the difference. The distance ratio provides good quantitative fits to some data sets, including data from experiments in which an *explicit* reference point is specified at an intermediate point on the scale (e.g., judging which digit, 2 or 3, is closer to 5; Holyoak, 1978). However, other data sets are less well fit by the quantitative form specified by the distance ratio. For example, although scale compression triggered by the form of the comparative can be observed in non-speeded discriminability judgments, the effects tend to be smaller than the distance ratio would predict (Holyoak & Mah, 1982).

Perhaps reference points directly alter mean magnitudes of items, expanding differences close to the reference point relative to differences far from it. However, shifts in discriminability might instead reflect changes in *variances* of magnitude, rather than in mean values. Marks (1972), building on the assumptions of signal detection theory, suggested that internal magnitudes are represented as distributions that encode uncertainty, which is reduced in the region of a reference point (i.e., the variance or "discriminal dispersion" of magnitude representations is lower for magnitudes close to a reference point). Marks did not develop a quantitative model; however, related reference-point models have introduced evidence-accrual mechanisms, consistent with the basic idea that comparative judgments are based on iterative sampling from magnitude distributions (see Petrusic, 1992). The model we propose in the present paper adopts the key idea proposed by Marks (1972), that the form of the

comparative affects discriminability by dynamically altering magnitude variances based on distance from a reference point.

Reference-point models in general, including Marks's (1972) specific proposal of the modulation of variance as a mechanism, are broadly consistent with the wider literature on attentional influences on magnitude representation. Miller's (1956) classic paper focused on the limited channel capacity available to make absolute magnitude judgments (and explicitly linked signal variance with information transmission). In psychophysical work, Luce, Green, and Weber (1976) proposed that observers are able to strategically control *attention bands*, selectively monitoring a relatively narrow intensity range. Luce et al. suggested that neural variability of the internal representation of intensities will be reduced within the favored attention band, yielding greater sensitivity as measured by signal-to-noise ratio. Nosofsky (1983) found evidence that observers can indeed strategically shift attention to a specific intensity band, thereby facilitating discrimination of tones in the favored region. He also argued, based on a literature review, that this flexible allocation of attention to a magnitude band is limited to just one such location along a continuum; hence performance falls off monotonically with distance from the favored region.

A reference-point explanation has also been offered for the markedness effect. It is possible that markedness, like the congruity effect, fundamentally arises from the inherent meaning of comparatives, and in particular from the fact that many comparative pairs have an inherent asymmetry in their polarity: one end is positive or "greater" and the other end is negative or "lesser". If markedness is rooted in the underlying meaning of comparatives, then the effect might reflect some additional processing difficulty encountered in maintaining precise magnitude distributions when focusing on the "negative" or "lesser" pole. Marks (1972) suggested that the markedness effect could be modeled by assuming that the precision of magnitude representations falls off more rapidly moving from the lesser than from the greater reference point. We will also adopt this assumption, which serves to integrate the markedness effect with the semantic congruity effect.

In sum, psychophysical work provides broad support for the hypothesis that observers can selectively modulate attention to a favored region along a magnitude continuum. Given the many established parallels between perceptual and symbolic magnitude comparisons, it is natural to hypothesize that similar mechanisms operate in symbolic tasks. Moreover, reference points established by the form of the question and the range of the presented stimuli can readily be viewed as cues that establish attention bands. Mark's (1972) proposal that such modulation operates by influencing the variance of magnitude representations provides a key theoretical element in the model we will describe below. The hypothesis that attention operates in part by modulating variability in an internal representation is also consistent with findings concerning visual detection and discrimination tasks (Dosher & Lu, 2000; Rahnev et al., 2011).

2. Magnitude representations in BARTlet

2.1. Multiple levels of representation for comparative relations

Our goal in the present paper is to provide a unified model of how symbolic magnitudes can be discovered and used to make comparative judgments. The model we propose, termed BARTlet (i.e., the diminutive form of BART), builds on the learning capability of BART (Lu et al., 2012) but makes simpler representational assumptions. A key idea incorporated in both models is that learning can be bootstrapped by incorporating *empirical priors*—a "favorable" initial knowledge state derived from some related but simpler learning task. In BART, learning of explicit comparative relations (two-place predicates, such as *larger*) is guided by empirical priors derived from initial learning of one-place predicates (e.g., *large*, *small*).

BARTlet also emphasizes the role of bootstrapping operations that allow learning at a lower level to guide subsequent learning at a higher level (for a more detailed discussion of bootstrapping, see Lu et al., 2012, p. 618). Although we do not aim to provide a serious developmental model (which would require a detailed specification of the inputs available to children), we do aim to implement a learning process that can acquire magnitude information from inputs of realistic complexity. Moreover, we

focus on learning from inputs that were not hand-coded, but rather were generated by an autonomous process (i.e., independently of the modelers). We use two different sets of inputs as a further way of showing that the learning model is robust and does not depend on specific details of what features are included in the input.

Given that humans are able to make magnitude comparisons between objects that they may never have previously considered together (e.g., which is larger, a walrus or a fox?), our goal was to create a model that can learn from a limited set of examples and then generalize to novel comparisons. At the same time, we also wished to capture the significant commonalities between magnitude comparisons performed by humans and by non-human animals. BART learns explicit two-place relations representing comparatives (e.g., *larger*). In addition to supporting generalization to new animal pairs, these explicit relations can be systematically transformed to solve analogies based on higher-order relations between different pairs of polar adjectives (e.g., *larger: smaller*:: *faster: slower*). However, such high-level reasoning is beyond the capability of most animals (indeed, it may be uniquely human; Penn, Holyoak, & Povinelli, 2008). In contrast, basic comparative judgment appears to be similar in humans and symbol-trained monkeys (Diester & Nieder, 2010; Moyer & Landauer, 1967). Many other species, such as rats, can respond on the basis of relative magnitude when shown perceptual stimuli that vary along simple continua (Lawrence & DeRivera, 1954). Thus as a model of basic comparative judgment, the explicit relational representations acquired by BART appear to be over-powerful.

Fig. 2 sketches different levels of representation that may be involved in making magnitude comparisons and reasoning with comparative relations (for a similar representational hierarchy, see Halford, Wilson, & Phillips, 1998, 2010). At a pre-categorical level (i.e., a level of representation that does not involve categorical distinctions or explicit predicates), simple associative or statistical mechanisms can perform basic magnitude comparison and learn from ordered pairs. For example, under certain conditions the Rescorla-Wagner model of associative learning (Rescorla & Wagner, 1972; see Wynne, 1995) can model qualitative aspects of animals' ability to infer transitivity of choice (e.g., after being trained on only adjacent pairs of stimuli exhibiting the reward pattern A > B, B > C, C > D, D > E, an animal will tend to choose B over D). Other associative models can account for learning of orderings across a broader range of conditions (von Fersen, Wynne, Delius, & Staddon, 1991).

In the present paper we adopt a statistical model capable of learning continuous-valued attributes from a *partial* ordering of examples (Parikh & Grauman, 2011). This model (described more fully below) learns to rank objects based on the algorithm of a support vector machine with certain additional

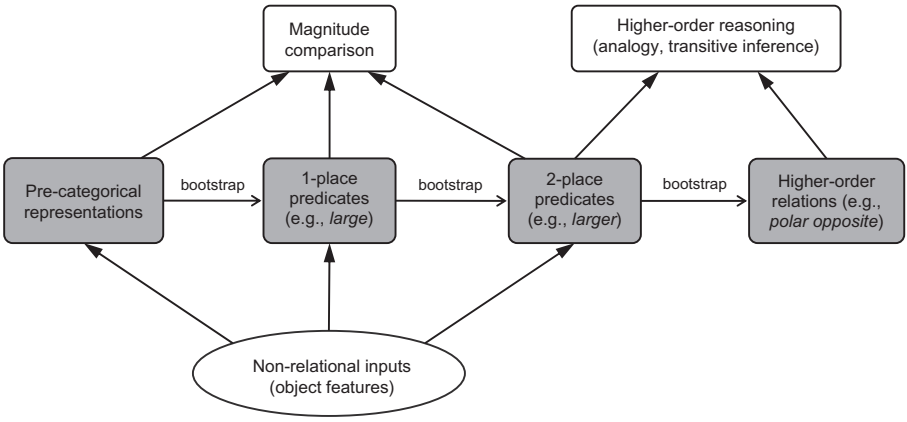

Fig. 2. Relationships among inputs (bottom), levels of representation (middle row) and tasks (top) involving magnitude-related concepts. Pre-categorical processes bootstrap acquisition of one-place predicates (the domain of BARTlet), which in turn can bootstrap acquisition of two-place predicates and ultimately higher-order relations (the domain of BART). The lower levels have access to external inputs (non-relational feature vectors for individual objects) and can be used to perform comparisons based on dimension-specific magnitudes; the higher levels operate (in part or entirely) on internally-generated representations, and can be used to perform more abstract types of reasoning, such as higher-order analogical reasoning and transitive inference.

constraints, and hence will be referred to as *RankSVM*. RankSVM extracts continuous dimensions of attributes by learning weights on object features, such that the maximum number of ranking constraints is satisfied for the training data. Note that RankSVM does not create representations that categorize attributes in a binary manner (e.g., elephant is large, not small); rather, this algorithm yields representations sensitive to relative order on a dimension (e.g., elephant is ordered before horse in size, horse is ordered before cat). Parikh and Grauman successfully tested their RankSVM model on problems involving comparisons of realistic visual images. Though we do not claim that the algorithm is psychologically realistic, it provides a functional model that can deal with partial orderings of elements coded by high-dimensional feature vectors. The function performed by this model is consistent with empirical evidence that both animals and humans can learn simple orderings from a partial set of ordered pairs (Merritt & Terrace, 2011; Trabasso & Riley, 1975; Woocher, Glass, & Holyoak, 1978; Wynne, 1995). Moreover, its output (feature weights) can readily be translated into empirical priors for learning one-place predicates.

The next level of representation corresponds to one-place predicates (e.g., *large*), which in essence define categories of objects based on their magnitudes on some underlying dimension. Both behavioral and neural evidence indicates that monkeys are capable of acquiring categorical representations (e.g., Cromer, Roy, & Miller, 2010; Freedman, Riesenhuber, Poggio, & Miller, 2001). For realistic stimulus sets, learners are unlikely to ever compare all possible pairs of N objects (a quantity that scales with N^2) on every dimension of interest. Categorical information about individual objects (a quantity that scales linearly with N) provides an efficient additional input for learning magnitudes. As described below, BARTlet learns one-place predicates from facts such as "a whale is large," bootstrapping from empirical priors provided by RankSVM. BARTlet thus integrates dimensional information provided by examples of ordered pairs (via RankSVM) with categorical information, thereby refining its knowledge about dimensional magnitudes.

The additional levels of representation sketched in Fig. 2 are based on explicit relations (i.e., predicates with more than one argument, such as *larger*). Whereas BARTlet uses a comparison operator (based on signal detection theory) to compare relative magnitudes derived from one-place predicates, BART creates two-place predicates that in effect represent the comparison operator as part of the relation itself. As described by Lu et al. (2012), these more complex relational representations (arguably unique to humans) can be learned by bootstrapping from one-place predicates, and can in turn be bootstrapped to generate higher-order relations between relations (e.g., "polar opposite"). We will return to the topic of levels of representation in the General Discussion. For now, we simply note that the goal of the present paper is to show that BARTlet, a model limited to one-place predicates (i.e., without access to explicit two-place comparatives) is capable of basic symbolic magnitude comparisons.

2.2. Deriving magnitudes from unstructured feature vectors

In BARTlet, magnitudes are created by applying learned dimension-specific weights to more primitive features of objects. Magnitudes are represented in working memory as derived features that follow specified probability distributions, modulated by reference points. BARTlet (like BART) represents a one-place predicate (e.g., *large*) using a joint distribution of weights over object features, as illustrated in Fig. 3 (bottom). A predicate is learned by estimating the probability distribution $P(\mathbf{w}|\mathbf{X_S}, \mathbf{\Phi_S})$, where $\mathbf{X_s}$ represents the feature vectors for objects in the training set, the subscript $\mathbf{S}$ indicates the set of training examples, and $\mathbf{\Phi_S}$ is a set of binary indicators, each of which (denoted by Φ) indicates whether a particular object instantiates the predicate or not. The vector $\mathbf{w}$ constitutes the learned predicate representation, which can be interpreted as weights reflecting the influence of the corresponding feature dimensions in $\mathbf{X}$ on judging whether the predicate applies. Formally, the posterior distribution of weights can be computed by applying Bayes' rule using the likelihood of the training data and the prior distribution for $\mathbf{w}$:

$$P(\mathbf{w}|\mathbf{X_S}, \mathbf{\Phi_S}) = \frac{P(\mathbf{\Phi_S}|\mathbf{w}, \mathbf{X_S})P(\mathbf{w})}{\int_{\mathbf{w}} P(\mathbf{\Phi_S}|\mathbf{w}, \mathbf{X_S})P(\mathbf{w})}. \quad (1)$$

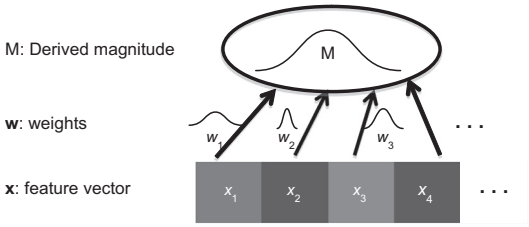

Fig. 3. In the BARTlet model, weight distributions derived from one-place predicates (e.g., *large*) are applied to the feature vector for an individual animal to compute a derived magnitude (normally distributed) for that object.

The likelihood is defined as a logistic function for computing the probability that an object instantiates the predicate, given the weights and feature vector:

$$P(\Phi = 1|\mathbf{w}, \mathbf{x}) = (1 + e^{-\mathbf{w}^{\mathbf{T}}\mathbf{x}})^{-1}. \tag{2}$$

The prior distribution $P(\mathbf{w})$ in Eq. (1) is assumed to follow a normal distribution with mean and covariance matrix as parameters. To define the prior, the BARTlet model relies on initial learning at a simpler representational level to bootstrap subsequent learning at a more complex level. Specifically, BARTlet uses weights learned by RankSVM as means and a standardized covariance matrix (e.g., variance of 1, covariance of 0) as the empirical prior for learning one-place predicates. The RankSVM model takes ordered pairs as inputs, where each object is represented by a feature vector. Its algorithm is a support vector machine, which in essence performs linear regression with an additional constraint to minimize weight values. The novel feature of RankSVM is the further addition of a penalty for violating the given partial ordering of objects (for a full mathematical description, see Parikh & Grauman, 2011).

RankSVM was developed for machine-learning purposes, and we make no claim for the psychological plausibility of its algorithm. However, there is ample evidence that many types of animals can learn simple orderings from a partial set of pairs. For both animals (Merritt & Terrace, 2011; Wynne, 1995) and humans (Trabasso & Riley, 1975; Woocher et al., 1978), orderings are typically learned "from the ends in", with the extreme or "landmark" objects being acquired prior to those that lie closer to the middle of a continuum. In the present simulations, we trained RankSVM with ordered pairs that mainly involved the half dozen animals with the highest or lowest values on the relevant continuum. The resulting weights then served as empirical priors for BARTlet, which in turn received relatively extreme animals as examples (positive or negative) of each one-place predicate.

2.3. From weight distributions to derived magnitudes

The weight distribution that BARTlet acquires for a one-place predicate such as *large* provides all the information required to specify the magnitude of each animal on each dimension. As shown in Fig. 3, the magnitude of an object on a dimension (e.g., size) can be derived as a weighted sum of the feature values $\mathbf{x}$ for this object:

$$M = \sum_i w_i x_i, \tag{3}$$

This weight distribution codes not only first-order statistics (means, μ_{w_i}), but also second-order statistics (variances and covariances) that capture the uncertainty of the estimated weights, as well as inter-weight correlations. Because the weights are normally distributed, the derived magnitude variable M follows a normal distribution with a mean of μ_M and a variance of σ_M^2, which are calculated according to:

$$\mu_M = \sum_i \mu_{w_i} x_i, \tag{4}$$

$$\sigma_M^2 = \sum_i x_i^2 Var(w_i) + \sum_i \sum_{j \neq i} x_i x_j Cov(w_i, w_j). \tag{5}$$

The variance of the derived magnitude reflects uncertainty about the magnitude value and can be modulated by factors such as attention, in a manner that we will describe. Importantly, BARTlet does not make use of explicit relations when making symbolic comparisons. Rather, BARTlet evaluates which of two objects is larger (or smaller, faster, etc.) by the more primitive operation of comparing the derived magnitudes of the two individual objects, using the framework of signal detection theory.

2.4. Reference points in symbolic comparisons

BARTlet adds two explicit algorithmic assumptions: People operate under limited capacity to maintain veridical estimates of magnitudes in working memory, and the focus of attention on a particular magnitude range is controlled by reference points. Because the representation of magnitudes includes uncertainty, it is straightforward to implement the key assumption that magnitude discriminability is influenced by reference points, which operate by influencing the associated variances (Marks, 1972). BARTlet selectively attends to a particular region of the relevant dimensional spectrum (e.g., the high end of the size spectrum when choosing the larger of two objects), leading to greater discriminability between objects in that favored region (Fig. 4). The distance to a reference point is calculated by comparing an object to a reference object, and this distance is used to scale the magnitude variance of the object. As a result, magnitudes of objects closer to the reference point have greater precision (i.e., less uncertainty), whereas the magnitudes of objects farther from the reference point have less precision.

BARTlet generates magnitude values (M) based on unmarked one-place predicates (e.g., *large*), and hence M values are positive and monotonic relative to the unmarked form (e.g., large animals are associated with high size values, and small animals with low size values, rather than the reverse). We assume that because the unmarked form of the question requires reversing the natural scale (e.g., "smaller" focuses attention on low magnitudes), precision diminishes more quickly with distance from the reference point in the case of the marked comparative.

Specifically, BARTlet uses the following procedure to answer a comparative query such as, "Which is larger, an elephant or a giraffe?" First, the model establishes a reference point based on the comparative involved in the question and all presented stimuli (i.e., the context). Because the comparative in this question is *larger*, the reference point is taken to be the object among the presented stimuli with

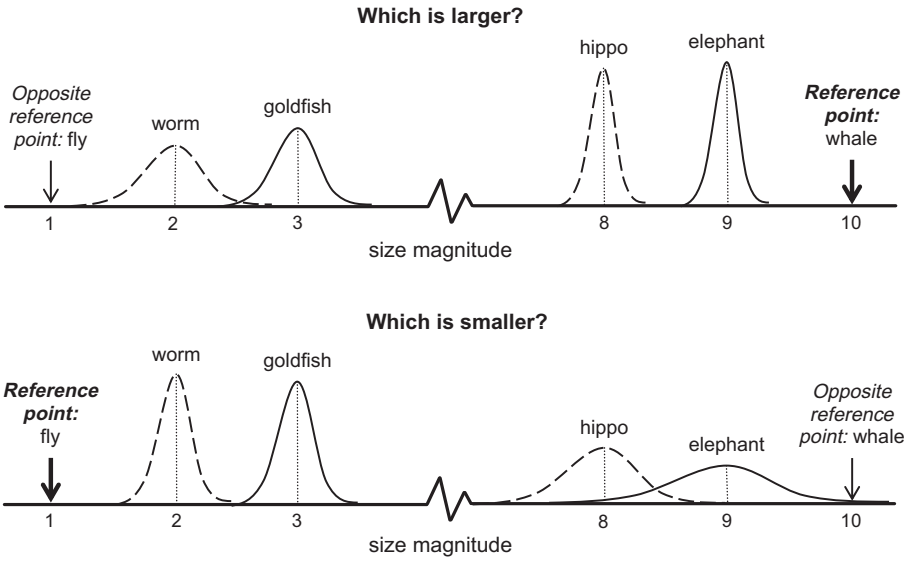

Fig. 4. BARTlet's representations of magnitudes in working memory. Based on the assumption that reference points at the extremes control attention, variances of magnitude distributions increase with distance from the reference point at the extreme consistent with the question. The increase in variance with distance from the reference point is assumed to be greater for the marked form of a comparative.

the highest mean magnitude on the size dimension. (If the comparative were instead *slower*, the reference point would be the object with the lowest mean magnitude on the speed dimension.) Based on the selected reference point, the model computes *D*, the maximum possible distance from the reference point within the current context (i.e., the subjective range on the relevant dimension). This value is simply the absolute difference in mean magnitudes between the reference point and the opposite-extreme reference point. For *larger*, the opposite-extreme reference point is the object among the presented stimuli with the lowest mean size magnitude.

The model computes the means and unscaled variances according to Eqs. (4) and (5) for the magnitudes of the two objects being compared. In our example, the mean and variance of the size magnitude is computed for both the elephant and the giraffe. Then, for each object being compared, the model computes δ, a measure of the distance between that object and the reference point as a proportion of the maximum possible distance from the reference point.[1] This value corresponds to the absolute difference between the mean magnitudes of the object and of the reference point, divided by *D*. For each object being compared, the model scales the variance of its magnitude by $\alpha e^{\beta\delta}$, where α is an intercept parameter and β is a slope parameter, both free parameters. The specific parameter values were selected to be consistent with the qualitative assumptions of the model. In our simulations, α was set to 0.1, implying that the variance of an object's magnitude is decreased by 90% when that object's mean magnitude is equal to that of the reference point. The values of β were selected so as to yield magnitude variances that are about 10 times (for unmarked relations; $\beta = 4.6$) or 20 times (for marked relations; $\beta = 5.3$) as high as the original variances when an object is maximally distant from the reference point. Thus, magnitude variances are assumed to increase more rapidly for marked relations than for unmarked relations as distance from the reference point increases (cf. Marks, 1972). In the present model, variances increase exponentially with distance from the reference point; however, a variety of neural mechanisms for gain control could potentially implement the impact of attention on gain control (Dosher & Lu, 2000; Rahnev et al., 2011; for a review see Reynolds & Chelazzi, 2004).

2.5. Measuring discriminability between magnitudes

BARTlet models the discriminability between magnitudes of two objects that are made available to a comparison process. Based on signal detection theory, a natural measure of discriminability is d_a, which is the variant of d' appropriate when variances are unequal (Wickens, 2002, p. 65):

$$d_a = \frac{\mu_{M_1} - \mu_{M_2}}{\sqrt{(\sigma^2_{M_1} + \sigma^2_{M_2})/2}}. \tag{6}$$

A complete model of symbolic magnitude comparisons needs to specify a decision process that would translate degree of discriminability into accuracy and reaction time for comparative judgments. For example, the decision diffusion model (Ratcliff, 1978; Ratcliff & McKoon, 2010; Ratcliff, Van Zandt, & McKoon, 1999) is an extension of signal detection theory to the time domain, accumulating information continuously on the basis of repeated samples (also see Link, 1990; Petrusic, 1992). The diffusion model has a plausible neural realization (e.g., Ratcliff, Cherian, & Segraves, 2003; Wong & Wang, 2006). If applied to comparative judgment, a theoretical measure of discriminability, such as d_a, could be used to predict the average value across repeated samples (corresponding to the mean of the drift rate in a diffusion process). Because our present focus is on variables that influence discriminability (i.e., information quality), rather than on the decision process *per se*, we will simply use BARTlet to make qualitative predictions of decision difficulty, based on values of d_a. We assume (as the diffusion model predicts) that decreases in discriminability will make the decision process more difficult, yielding slower and/or less accurate comparative judgments.

[1] We assume for simplicity that reference points are established using the range of presented stimuli. Of course, the range of presented stimuli will typically become apparent to the observer over the course of exposure to a series of examples. Reference points are therefore likely to be updated dynamically, reflecting a compromise between prior expectations about stimulus range and the range actually observed in the context (Petrusic & Baranski, 1989).

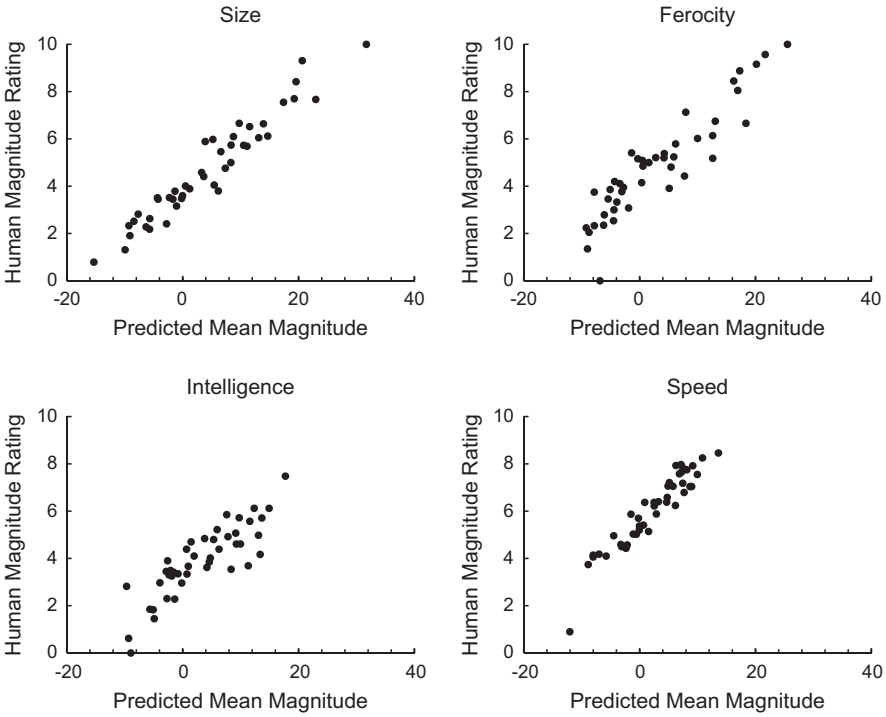

Fig. 5. Scatter plots of human magnitude ratings (based on data from Holyoak and Mah, 1981) versus mean magnitudes derived from BARTlet using Leuven vectors for animals on four dimensions.

3. Simulations of symbolic magnitude judgments using Leuven vectors

3.1. Predicting human magnitude ratings

We first evaluated whether the *M* values learned by BARTlet in fact reflect the subjective magnitudes of animals on the relevant dimensions. The "ground truth" for all training examples and test pairs was provided by norms derived from ratings by college students on the dimensions of size, ferocity, intelligence and speed (Holyoak & Mah, 1981). For the animals used in the simulations reported in the present paper, intercorrelations among the four dimensions were moderate, ranging from .38 (size with speed) to .60 (size with fierceness). For our first set of simulations, we identified a set of 44 animals that also appeared in the Leuven norms (de Deyne et al., 2008). Each animal was represented by a vector of 50 continuous-valued features (see Lu et al., 2012, pp. 631–632, for a description of how the Leuven vectors were created).

As described earlier, learning of one-place dimensional predicates (*large*, *fierce*, *intelligent*, *fast*) proceeded in two stages. First, RankSVM was provided with the ordering for each of the top three and bottom three animals on the relevant dimension relative to all other animals, intermixed with an additional 100 pairwise orderings selected at random from the pool of all possible pairs of 44 animals.[2] The mean weights estimated by RankSVM (linearly scaled by a factor of 5 to roughly match the range of weights BARTlet would infer from an uninformative prior) became the empirical priors

[2] The specific selection of training examples is not critical to the performance of the model. We aimed to limit the number of training examples so that the model was forced to generalize on test pairs. The emphasis on early learning of extreme "landmark" animals is consistent with the typical pattern observed in learning orderings (Potts, 1974; Rylass & Smith, 2000).

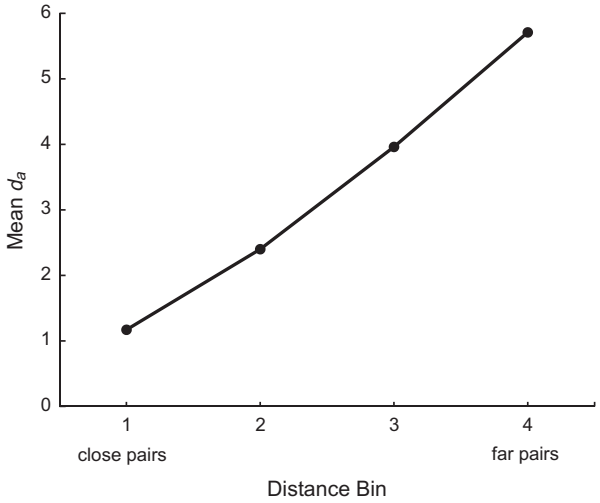

Fig. 6. BARTlet's predicted discriminability value, d_a, for comparative judgments using Leuven vectors as a function of the subjective distance between pairs of animals on the relevant dimension. Distance bins are based on Holyoak and Mah's (1981) norms, in which values range from 0 to 10: bin 1 (distances between 0.5 and 1.5), bin 2 (distances 1.5–3), bin 3 (distances 3–5.5), and bin 4 (distances 5.5–10). Results are collapsed over the four unmarked comparative relations.

on weight means for BARTlet.[3] As RankSVM does not provide a covariance matrix, an uninformative prior (variances = 1, covariances = 0) was used. Second, BARTlet was provided with the 20 animals with the highest values (positive examples) and the 20 with the lowest values (negative examples) on the relevant dimension. These training examples were drawn from the entire pool of 129 animals in the Leuven norms. The resulting weight distributions across the 50 features of the Leuven inputs (Fig. 1) were highly distributed, based on at least 20 statistically predictive features for each of the four magnitude dimensions of interest.

The weight distribution for each one-place predicate was used to calculate *M* values for each animal, as described earlier. Fig. 5 shows the scatter plots of mean *M* values versus human magnitude ratings for each of these dimensions. Spearman rank-order correlations ranged from .86 to .96 for the four dimensions. These results indicate that magnitude values, derived from weight distributions acquired by BARTlet's learning mechanism from large, independently-generated feature vectors (Leuven vectors; see Fig. 1), are quite accurate in predicting human judgments about subjective magnitudes of animals on the four dimensions.

3.2. Symbolic distance effect

To evaluate whether BARTlet exhibits the ubiquitous symbolic distance effect obtained for comparative judgments by humans, we formed all possible pairs of the 44 animals previously identified, which served as testing items for each of the unmarked comparative relations corresponding to the four rated dimensions in Holyoak and Mah's (1981) norms: *larger*, *fiercer*, *smarter*, and *faster*. To ensure that the differences in magnitudes between animals in a pair were likely to be distinguishable by humans, we excluded pairs that differed by less than .5 on the normed ratings for the relevant dimension. The resulting pairs of animals were grouped into four distance bins, such that animals very close on the relevant dimension fell into bin 1 and animals maximally far apart on that dimension fell into bin 4. Fig. 6 plots the mean d_a value for each distance bin, averaged across the four unmarked compar-

[3] The use of the prior provided by RankSVM increased the rank-order correlations between human magnitude ratings and magnitudes derived from the model by approximately .10 (relative to an uninformative prior) for the Leuven inputs and about .02 for the topics inputs.

ative relations. Results for the four marked relations are similar. Consistent with a symbolic distance effect, BARTlet's predicted discriminability increases with the distance between the pair of animals.

3.3. Semantic congruity effect

To test BARTlet's ability to predict the congruity effect, for each of the four dimensions we selected five animal pairs that were either both at the high end (e.g., *whale–elephant* for size) or both at the low end (e.g., *goldfish–fly*). We selected pairs that were at least minimally discriminable based on the learned weight distributions. All these pairs were relatively close in magnitude, as the congruity effect is typically maximized when both pairs are near to an extreme and hence close in magnitude. A congruity effect was observed for all four dimensions, as indicated by the interaction apparent in each panel (see Fig. 7). In each case the interaction shows an asymmetry, with the advantage of the unmarked congruent form of the question (e.g., "choose larger" for large animals) being slightly greater than the corresponding advantage of the marked congruent form (e.g., "choose smaller" for small animals). In other words, the congruity effect was modulated by a markedness effect, as is commonly observed in behavioral studies (e.g., Holyoak & Mah, 1981).

3.4. Influence of stimulus range on congruity effect

An important additional finding concerning the congruity effect is that it is influenced by the range of magnitudes represented in the stimulus set (e.g., Čech & Shoben, 1985, for humans; Jones et al., 2010, for monkeys). Since BARTlet sets its reference points dynamically based on the magnitude range relevant to the current context, it naturally predicts how the congruity effect will vary with the context. To test this aspect of the model, we created four sets of stimuli based on the size dimension, or-

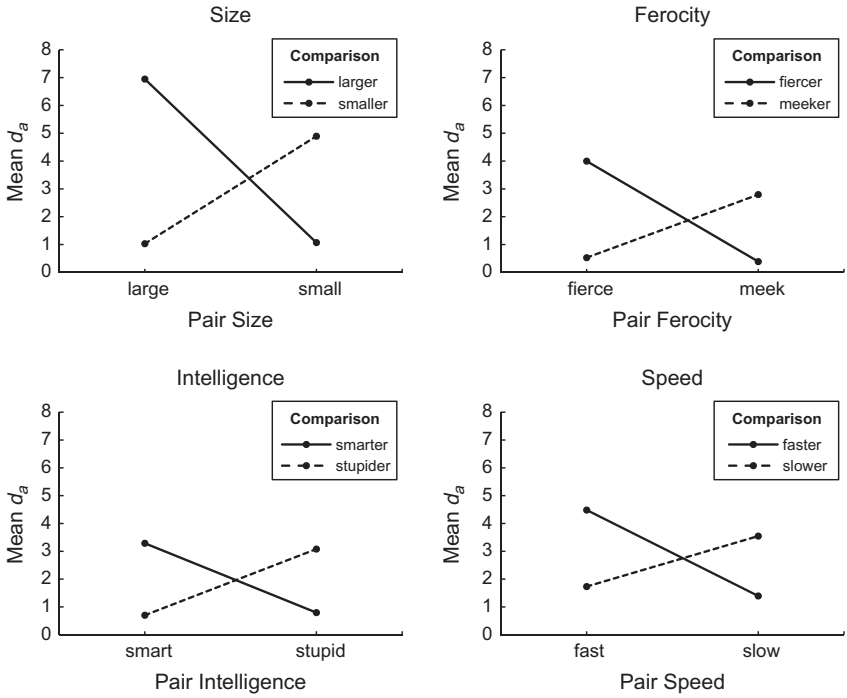

Fig. 7. Predicted semantic congruity effect for magnitude comparisons with polar adjectives using Leuven vectors, based on BARTlet's predicted discriminability value, d_a, for unmarked and marked comparatives for four dimensions.

dered in size from Set 1 (pairs of largest animals) to Set 4 (pairs of smallest animals). Sets 1 and 4 were the same pairs used to test the basic congruity effect (see Fig. 7). Sets 2 and 3 were intermediate in size (e.g., Set 2 included *alligator-pig*; Set 3 included *cat-sparrow*). The size distance between the two animals in each pair was closely matched across all four sets. In two different tests, BARTlet made "choose larger" and "choose smaller" judgments using either the full range of magnitudes (i.e., Sets 1–4), or a restricted range (i.e., Sets 2–3 only). As shown in Fig. 8, both tests yielded congruity effects; however, the magnitude of the congruity effect for the critical Sets 2–3 based on middle-sized animals was substantially larger when these intermediate sets were tested alone (restricted range; 2.03 in d_a units) than when they were intermixed with the pairs of very large or very small animals (full range; 1.11 in d_a units). BARTlet thus provides an account of how context can influence comparative judgments by dynamically altering reference points.

4. Simulations of symbolic magnitude judgments using topics vectors

To derive topics vectors, we obtained a preprocessed version of the English Wikipedia corpus in which entries shorter than 512 words were removed, as were words that are not in a standard English dictionary or that are on a list of "stop words" (high-frequency function words that have low semantic content, such as *the*, *and*, etc.), resulting in a total of 174,792 entries and 116,128 unique words. We ran the topic model (Griffiths et al., 2007) on this corpus to obtain 300 topics. The algorithm was used to generate three Markov chains, taking the first sample after 1000 iterations and then sampling once every 100 iterations, for a total of eight samples from each chain or 24 samples overall. Each sample yielded a matrix in which the (i, j)th entry is the number of times that word i has been assigned to topic j. From this matrix, we derived a vector for each word based on the conditional probability of each topic given that word (i.e., each resulting word vector is based on the relative frequencies of the different contexts within which the word could occur), using the same procedure employed by Lu et al. (2012) for outputs of the topic model ran on a different corpus.

Samples from a single Markov chain were very similar, in that the same 300 topics seemed to be found in each (based on examining the most probable words for each topic), but different chains produced different sets of topics. To create a single unified set of topics vectors for all words, we first averaged the word vectors based on samples from the same chain to produce a single set of word vectors for each chain. We then unified the three different chains (averaged across eight samples each) through the following procedure: First, for each of the averaged chains, we chose the 30 features

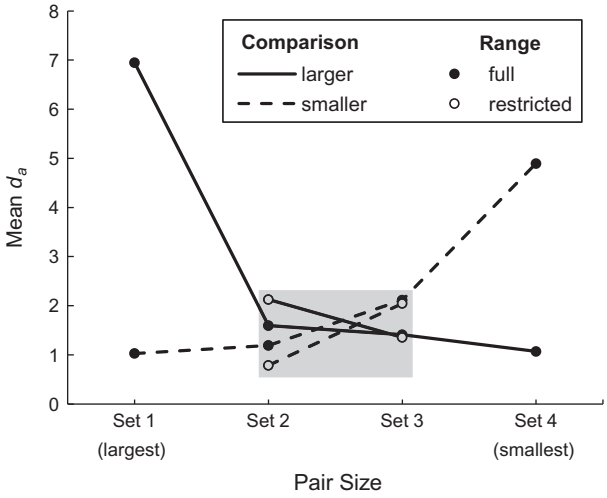

Fig. 8. Predicted semantic congruity effect (using Leuven vectors) for stimuli from the full range of animal sizes (Sets 1–4) and a restricted intermediate range (Sets 2–3 only), based on BARTlet's predicted discriminability value, d_a.

(topics) that had the highest sums across the vectors of the 77 animal words in Holyoak and Mah's (1981) norms (i.e., the 30 most prevalent topics for these animal concepts). Using the resulting animal vectors (reduced to 30 features for each chain), we then ran the full BART model to learn the relations *larger*, *fiercer*, *smarter*, and *faster*. We examined BART's generalization performance for these relations using the animal vectors from each chain (using the same tests as Lu et al., 2012). Starting with all 30 features from the chain that produced the best performance, we added features one at a time from the other two chains (each of which also had 30 features) in order of BART's performance on the chains. To minimize redundancy, a feature was added only if its correlations across the 77 animals with the features chosen so far were all less than .80. This process resulted in a total of 52 selected features. All simulations reported below were run using these topics vectors of length 52.

Based on the topics vectors, the same general procedure was used to learn one-place predicates with BARTlet as was used for Leuven vectors (i.e., initial weights acquired using RankSVM provided empirical priors for the learning of one-place predicates). The weights obtained by RankSVM were scaled by a factor of 10 rather than by a factor of 5 (to better match the scale of weights learned from topics vectors). The top and bottom 20 animals on each dimension (used as training data for BARTlet after the RankSVM stage) were drawn from the 77 animals in Holyoak and Mah's norms, rather the 129 animals in the Leuven dataset. The same method was used as before for calculating magnitude means and variances for each animal on each dimension.

4.1. Predicting human magnitude ratings

As we had done for the Leuven vectors, we performed correlational analyses to predict the human ratings (from Holyoak & Mah, 1981) using magnitudes extracted from the one-place predicates learned by applying BARTlet to topics vectors (except across a total of 77 animals, rather than the

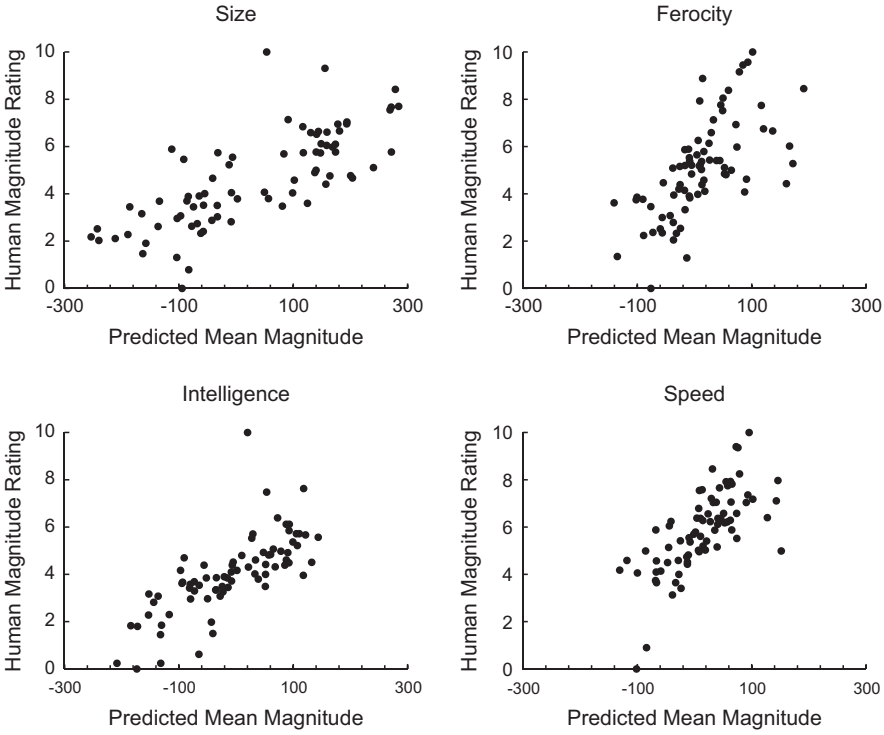

Fig. 9. Scatter plots of human magnitude ratings (based on data from Holyoak and Mah, 1981) versus mean magnitudes derived from BARTlet using topics vectors for animals on four dimensions.

44 available with Leuven vectors). Scatter plots are shown in Fig. 9. Spearman rank-order correlations were lower than for the Leuven vectors, but all were reliable, ranging from .73 to .82 across the four dimensions.

4.2. Symbolic distance effect

As shown in Fig. 10, the topics vectors yielded a robust distance effect (calculated in the same way as for the Leuven vectors, except using an additional distance bin made possible because a larger set of animals was available).

4.3. Semantic congruity effect

As done previously for Leuven vectors, we selected sets of five pairs of animals consisting of animals near the high or else low end of each of the four continua. Each pair was at least minimally discriminable but relatively close in magnitude (as the congruity effect is maximized for comparison of items with similar magnitudes).

Fig. 11 shows the congruity effects obtained for each of the four dimensions. A robust congruity effect was obtained for each. A markedness effect (overall advantage for the unmarked form of the comparative) was obtained for all of the dimensions (though more pronounced for size and ferocity than for the other two). Because the topics vectors yielded cruder magnitude codes than did the Leuven vectors, we did not attempt to model the effect of range (as it was too difficult to generate discriminable pairs at more than two levels of overall magnitude).

5. General discussion

5.1. Relational comparisons without explicit relations

In the present paper we have presented a model, BARTlet, that provides a unified account of how subjective magnitudes on different dimensions can be learned from more elementary features, represented and modulated in working memory, and used to assess the discriminability of objects. Previous models of symbolic magnitude comparisons have tacitly assumed that magnitude values on the

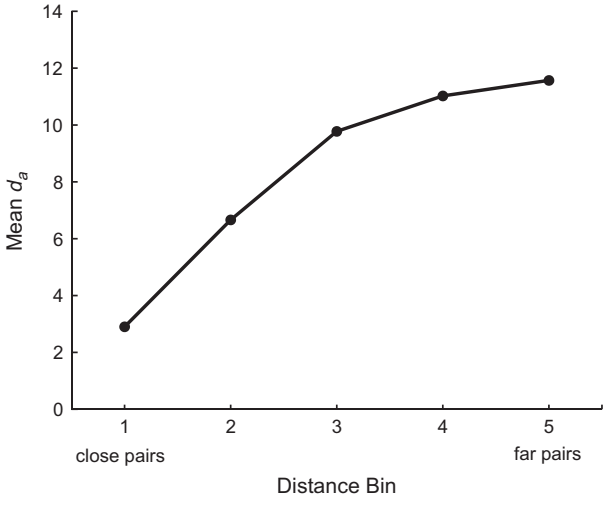

Fig. 10. BARTlet's predicted discriminability value, d_a, for comparative judgments using topics vectors as a function of the subjective distance between pairs of animals on the relevant dimension. Distance bins are based on Holyoak and Mah's (1981) norms, in which values range from 0 to 10: bin 1 (distances between 0.5 and 2), bin 2 (distances 2–4), bin 3 (distances 4–6), bin 4 (distances 6–8), and bin 5 (distances 8–10). Results are collapsed over the four unmarked comparative relations.

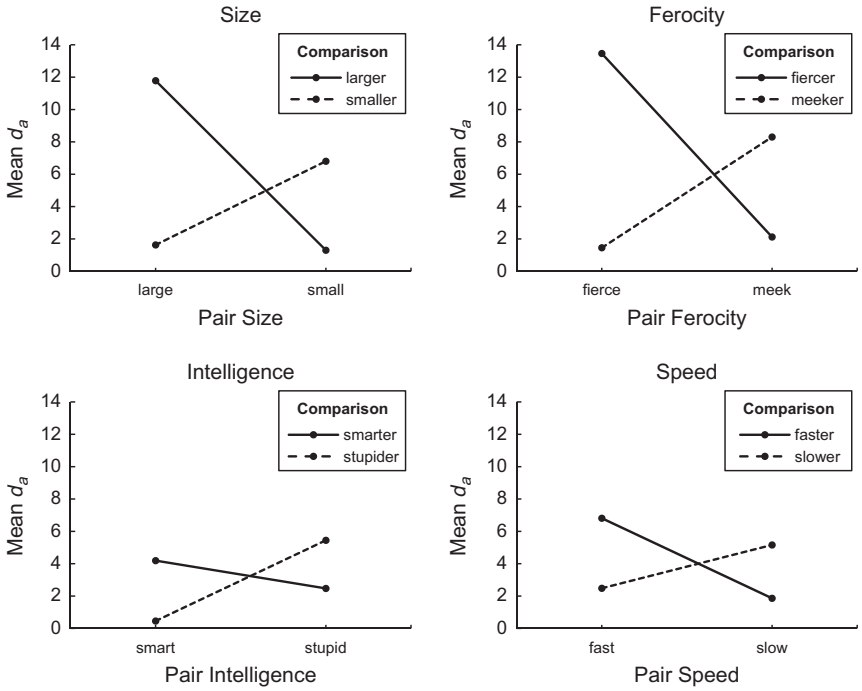

Fig. 11. Predicted semantic congruity effect for magnitude comparisons with polar adjectives using topics vectors, based on BARTlet's predicted discriminability value, d_a, for unmarked and marked comparatives for four dimensions.

relevant dimensions are prestored in long-term memory as features of objects. We argue that this assumption is unrealistic, even for a quasi-perceptual dimension such as size, and especially for the many complex social and interpersonal dimensions on which people can make comparisons (e.g., intelligence, religiosity). By building on BART, a Bayesian model of how comparative concepts can be learned from examples by statistical inference (Lu et al., 2012), we were able to integrate an account of how magnitudes are compared with an account of how magnitudes can be created in working memory based on prior learning about comparative concepts. The generality of the approach was demonstrated by applying the model to two sets of inputs (Leuven vectors and topics vectors), each of which was generated autonomously. BARTlet serves as an existence proof that symbolic comparisons can be modeled using high-dimensional distributed representations of elementary features, without assuming pre-existing dimensions, and without hand-coding inputs.

The operation of BARTlet, in comparison to its "smarter" precursor, BART, provides an instructive computational example of how a relational task (comparative judgment of magnitudes) can be performed without explicit relational representations. BART forms explicit representations of first-order relations such as *larger* (defined by weight distributions over pairs of objects assigned to distinct roles). In contrast, BARTlet operates only on weight distributions for one-place predicates (e.g., *large*), bootstrapping from priors on mean weights derived from pre-categorical comparisons (a partial ordering of pairs from which mean weights are learned by RankSVM, a model based on statistical regression). Magnitudes of individual objects are derived directly from the learned weight distributions for one-place predicates. BARTlet then proceeds to use an implicit comparison operation, which can be characterized in terms of signal detection theory, to assess which of two objects is the larger. No explicit *larger* relation is needed for BARTlet to choose the larger of two objects. BARTlet is thus an existence proof that the ability to make comparative judgments does not require explicit relational representations, consistent with evidence that rudimentary types of symbolic magnitude comparisons are within the capabilities of non-human primates (Cantlon et al., 2009).

Whereas BART is a computational-level model (Marr, 1982) of how comparatives can be learned, BARTlet adds explicit algorithmic assumptions concerning the representation and processing of magnitudes, based on consideration of limited computational resources in working memory. These core assumptions are firmly rooted in long-standing theories concerning attentional influences on magnitude representation. Human (and non-human) observers have limited capacity in working memory to maintain veridical estimates of magnitudes, which therefore vary in their precision (Miller, 1956). To partially compensate, observers focus attention on a favored region, or magnitude band, along the relevant continuum (Luce et al., 1976; Nosofsky, 1983). When making comparisons based on relative concepts, such as "choose larger" or "choose smaller", attention is guided by a reference point located at or near the end of the continuum cued by the form of the question (Holyoak, 1978; Jamieson & Petrusic, 1975; Marks, 1972). More specifically, selective attention causes the precision of magnitudes in working memory to be greatest (i.e., associated with low variance) for values close to the reference point, decreasing with distance from the reference value (Marks, 1972). The decrease in precision with distance from the reference point tends to be asymmetrical, with a steeper function for the "marked" form of the question (e.g., "choose fiercer" as opposed to "choose meeker").

Armed with these algorithmic assumptions, together with the tools of signal detection theory, we showed by a series of simulations that BARTlet can predict (1) human ratings of subjective magnitudes for animals along four different dimensions, (2) the symbolic distance effect, (3) the semantic congruity effect, (4) the modulation of the congruity effect by the polarity of the comparative (i.e., markedness), and (5) the context sensitivity of the congruity effect (i.e., the influence of the magnitude range of the presented stimuli). Furthermore, BARTlet accounts for all of these phenomena based on magnitude distributions that emerge from prior statistical learning of weight distributions over a high-dimensional feature space. No previous theory of magnitude comparisons has provided a comparable integration with the acquisition of comparative concepts.

5.2. Reference points in magnitude comparisons

BARTlet provides a computational realization of a qualitative hypothesis proposed four decades ago by Marks (1972): Reference points cued by the form of comparative questions systematically modulate the precision of magnitudes represented in working memory, yielding the semantic congruity effect. The reference-point hypothesis implies that the congruity effect results from differences in the discriminability of magnitudes represented in working memory, rather than a bias in encoding (e.g., Marschark & Paivio, 1979) or a linguistic influence (Banks et al., 1975). BARTlet provides a well-specified mechanism by which reference points can alter discriminability in direct judgments of discriminability (Holyoak & Mah, 1982) as well as speeded tasks. The modulation of precision will maximally impact discriminability between objects with relatively similar magnitudes, in accord with the general finding that congruity effects are larger when the objects being compared are closer in magnitude (Petrusic, 1992). The BARTlet model could easily be extended to account for the impact of explicit reference points (e.g., in a task requiring selection of which of two digits is closer in magnitude to 5; Holyoak, 1978), which can shift the favored attention band to an intermediate region on a continuum.

BARTlet generates magnitude values (M) based on unmarked one-place predicates (e.g., *large*), and hence M values are positive and monotonic relative to the unmarked form (e.g., large animals are associated with high size values, and small animals with low size values, rather than the reverse). We assume that because the unmarked form of the question requires reversing the natural scale (e.g., "smaller" focuses attention on low magnitudes), precision diminishes more quickly with distance from the reference point in the case of the marked comparative. Our approach thus provides a mechanism by which polarity could impact magnitude judgments made by non-linguistic animals (Cantlon & Brannon, 2005). This interpretation supports the hypothesis that the linguistic differences associated with markedness in human languages can be traced to more fundamental representational differences in magnitude continua.

The strong evidence that reference points influence discriminability implies that the semantic congruity effect is properly viewed as an example of the broader class of framing effects that impact decision making (Tversky & Kahneman, 1981). Indeed, semantic congruity effects have been observed

not only in magnitude comparisons involving objects, but also in judgments of preferential choice. For example, Birnbaum and Jou (1990) found that judging which individual is "liked more" for generally likeable individuals took less time than judging between unlikeable individuals, whereas judging which individual is "liked less" for likeable individuals took more time than judging between unlikeable individuals (also Nagpal & Krishnamurthy, 2008). The mechanisms instantiated in the BARTlet model may well prove applicable to decision making in areas such as consumer choice and social judgment.

The general notion of reference points has also been introduced in linguistic models of the interpretation of scalar adjectives (Tribushinina, 2009), which are interpreted in a context-sensitive manner. Scalar adjectives such as *large*, *warm*, and *average* refer to positions along a continuous dimension of magnitude; they are interpreted not as absolute values, but rather in relation to the noun category being modified (Partee, 1995). Thus an eagle is a large bird, but not an especially large animal; a tall boy is tall for a boy, but not for a tree. The interpretation of scalar adjectives requires scaling a subjective magnitude, or a probability of category membership derived from a magnitude, based on comparison to a norm or range derived from knowledge about the noun concept (see Barner & Snedeker, 2008).

5.3. The power and limits of magnitude representations

The parallels between the patterns of performance observed in monkeys and humans when performing magnitude comparisons suggest that this type of comparative judgment is based on evolutionarily primitive mechanisms. More broadly, neural and other evidence indicates that primates have evolved a specialized system for processing approximate magnitude, in which the intraparietal sulcus plays a key role (e.g., Cantlon et al., 2006; Dehaene & Changeux, 1993; Piazza et al., 2004, 2006, 2007; Pinel et al., 2004).

One reason for the apparent ubiquity of magnitude representations is that they can serve to answer multiple types of questions, each of which also provides learning opportunities. BARTlet learns magnitudes by integrating training with partial orderings (e.g., elephant is ordered before dog on the size dimension), the type of information provided to RankSVM, with training based on categorical inputs (e.g., elephant is large). Its acquired magnitude information might then be used to answer other interrelated types of questions (e.g., How large is a dog? Is a dog large? Is it larger than a cat? Is it smaller than a bear? Which is closer in size to a bear, a dog or a fox?). Feedback on the answers to any of such questions could be used to refine magnitude representations for a wide range of individual animals (not just those directly queried), thereby improving the model's ability to answer any question that depends on these magnitudes.

The fact that magnitudes are involved in answering many different questions and can be learned by multiple routes explains why evolution has apparently placed a premium on the creation of specialized neural hardware for manipulating such representations. Given the ubiquitous importance of comparative judgments in decision making, a system for discovering and manipulating magnitudes will be broadly advantageous. Nonetheless, unidimensional magnitude representations have their limitations. One limitation is that the neural system for approximate magnitude acts as a bottleneck. Precisely because any dimension can be coded in terms of a single internal number line, it is very difficult to code distinct orderings on separate dimensions for a single set of objects (Banks & White, 1982), a bottleneck that contributes to the "halo effect" (DeSoto, 1961). In addition, the validity of a one-dimensional magnitude representation is inherently limited, as is apparent whenever we try to reduce a complex multidimensional situation to a single number that serves as a "score" (e.g., GPA as a summary of a student's academic ability, *h*-index as a summary of a scientist's scholarly impact, dollar earnings as a summary of a year of one's life).

5.4. Limitations and possible extensions of the BARTlet model

Although the BARTlet model captures several basic phenomena related to symbolic magnitude comparisons, it currently has a number of empirical limitations. We have focused on the distance, semantic congruity, and markedness effects, which are arguably the phenomena most universally ob-

served in studies of symbolic magnitude comparisons. An additional phenomenon, typically observed for comparisons involving a closed-set series for which only ordinal information is available (e.g., an arbitrary ordering of elements for which magnitude information is not provided), is a bow-shaped serial position curve: accuracy and decision time indicate greater difficulty for pairs drawn from near the center of the list than for pairs closer to the ends. A bowed serial position curve is not observed for magnitude continua such as those on which we have focused in the present paper, but it is found for arbitrary orderings, both for humans (e.g., Potts, 1974; Trabasso & Riley, 1975; Woocher et al., 1978) and many animal species, including squirrel monkeys (McGonigle & Chalmers, 1977), rats (Davis (1992) and pigeons (von Fersen et al., 1991; for a review see Merritt & Terrace, 2011).

Although BARTlet does not currently model learning and performance with arbitrary series, it is in fact well-suited to be extended in this direction. One leading hypothesis is that bow-shaped serial position curves reflect *positional discriminability* (Holyoak & Patterson, 1981; Merritt & Terrace, 2011). The basic idea is that if individual items lack featural information that conveys magnitude, they are instead coded by their position relative to the beginning and end terms, which are learned first and serve as anchors. In accord with the representations used by BARTlet, these positional codes will be imprecise, approximating a normal distribution centered on an item's veridical position. Positional codes can be compared in the same way as codes for "true" magnitudes. The codes for central items will necessarily have greater overlap, and may well have higher variances than end items (Bower, 1971; Murdock, 1960; Trabasso & Riley, 1975). Thus, a natural extension of BARTlet would use the same basic type of representation—continuous-valued codes, normally distributed and varying in precision—to explain comparisons based on arbitrary ordered sets of elements. Such an extension would generate responses that exhibit distance effects, congruity effects, bow-shaped serial position effects, and transitivity of choice, as is empirically observed.

There has been some debate concerning whether, or in what way, magnitude codes are spatial in nature. The apparent empirical differences between learning and performance with dimensional magnitude codes versus positional codes suggest that although both are essentially analog (i.e., continuous-valued), magnitude codes are not necessarily spatial (nor are they inherently visual; Holyoak, 1977). In contrast, positional codes seem to be more spatial in nature, akin to an internal array (Holyoak & Patterson, 1981; Woocher et al., 1978). Nonetheless, similar brain areas are involved in comparisons of both types (see Cantlon et al., 2009).

A behavioral phenomenon often cited in support of a specifically spatial interpretation of magnitude codes, especially for number, is the SNARC effect ("Spatial Numerical Association of Response Codes"; Dehaene, 1992; Dehaene, Bossini, & Giraux, 1993). When evaluating a number (e.g., deciding whether it is odd or even), people typically respond to small numbers more quickly when the response key is to the left, and to large numbers more quickly when the response key is to the right. The SNARC effect thus suggests that number magnitude has a natural mapping onto the left–right axis of space (small numbers associated with the left).

The original tasks that exhibited a SNARC effect only used numbers, and did not involve magnitude comparisons. More recently, SNARC-like effects have also been observed in comparison judgment tasks, but the empirical picture is quite complex. Shaki, Petrusic, and Leth-Steensen (2012) reported that (1) a typical SNARC effect is found for digit comparisons with both "larger" and "smaller" instructions, (2) a typical SNARC effect is found for animal size comparisons with a "choose smaller" instruction, but a *reverse* SNARC effect is found for a "choose larger" instruction; (3) a short, newly-learned height ordering behaves much like size comparisons; (4) the above pattern for English speakers (1–3) is reversed for Israeli-Palestinians who habitually read right-to-left. A rough characterization of Shaki et al.'s (2012) findings is that although by default small numerical magnitudes are associated with the left, for non-numerical continua this bias is overridden by a preference to place the *reference point* on the left (or more generally, on the side from which orderings usually begin—hence the reversal due to cultural experience).

BARTlet does not model output processes, so it does not provide any obvious insight into the SNARC effect. However, as Shaki et al. (2012) noted, "...the mere fact that spatial information is being activated in association with the activation of magnitude information does not, in and of itself, conclusively imply that such spatial information is then actually being used by the comparison process

itself" (p. 525). Whatever the SNARC effect may imply about spatial processing, there is reason to doubt it has a deep connection to the comparison process that is the focus of BARTlet.

A further limitation of BARTlet stems from the fact that it can only compute comparative relations, and does not store or retrieve facts. People can certainly learn specific relational facts that arise repeatedly, or are tied to the intrinsic meanings of words (e.g., we commonly see dogs that are larger than cats; we know mountains are larger than hills because of how these terms are defined), and comparisons of this sort are made relatively quickly (Holyoak, Dumais, & Moyer, 1979). BARTlet does not account for the role of fact retrieval in magnitude comparison. It should be emphasized, however, that fact retrieval seems to play a modest secondary role. The initial demonstration of distance effects involving the digits 1–9 (Moyer & Landauer, 1967) was especially compelling because although adults surely know the fact that 3 is larger than 2 very well, they nonetheless find it easier to decide that 8 is larger than 2. In general, the ease of mental comparison seems to trump that of fact retrieval.

5.5. Relation to previous models of learning dimensional representations

As a learning model, BARTlet is based on the BART model, which Lu et al. (2012, pp. 640–642) discussed in relation to other models of relation learning. Here we consider three models (roughly ordered from least to most explicit in their relational representations) that have addressed the acquisition of continuous dimensions and/or linear orderings.

Smith, Gasser, and Sandhofer (1997) developed a multi-layer neural network model that learns dimensional adjectives by back-propagation. This model focuses on the interactive constraints provided by sensory, perceptual and linguistic information. Smith et al. argued that dimensional attributes, such as *large* or *red*, need not correspond to invariant features at the sensory level, but rather can be learned as distributed representations over more elementary features. Learning in their model involves updating weights on features; the magnitudes of weights are interpreted as indicators of learned selective attention. These assumptions are shared by BARTlet. Though the Smith et al. model has not been directly applied to the task of magnitude comparisons, it might well be extended in that direction. As a standard neural network, the model learns weights as point estimates, and hence does not capture differences in precision. But at a global level, the Smith et al. model is similar in spirit to BARTlet, taking a basically bottom-up approach to the acquisition of dimensional concepts, and operating without explicit representations of comparative relations.

DORA (Discovery of Relations by Analogy) is a symbolic-connectionist model that learns both one-place predicates (e.g., *large*) and two-place relations (e.g., *larger*), focusing on comparatives (Doumas, Hummel, & Sandhofer, 2008). Like BARTlet (and BART), it emphasizes bottom-up learning from objects coded as feature vectors (though it has not yet been tested on high-dimensional inputs of the sort used in the present paper). DORA includes a comparator operator that is well-suited for performing magnitude comparisons. Because DORA's predicates are initially most similar to the specific cases from which they were learned, the model predicts a congruity effect early in learning (e.g., for children, the representation of *large* will be more similar to large than small objects, and vice versa for *small*, leading to a congruity effect). As the model continues to refine its predicates using a feature-intersection mechanism, its representations of dimensional adjectives will tend to become more "magnitude neutral." It is therefore less clear whether the model could account for congruity effects observed in studies with adults. However, it is possible that DORA could be extended to include assumptions about the role of reference points.

Finally, an extremely general framework for learning relational structures has been proposed by Kemp and Tenenbaum (2008, 2009). By coupling a generative grammar for structural forms with a hierarchical Bayesian inference engine, their integrated model can generate many different structures to explain data patterns, including trees, multidimensional spaces, grids, rings, chains and (most importantly in the present context) linear orders. As Kemp and Tenenbaum acknowledge, "...we offer a modeling framework rather than a single model of induction. Our framework can be used to construct many specific models..." (2009, p. 22). Any specific model within the framework involves a combination of assumptions about the available forms and about the processes that operate on forms to make inductive inferences. Given its flexibility, a model could presumably be created within the framework that would closely emulate BARTlet (or BART, or other alternative models).

The power of the framework is also its Achilles' heel as a psychological theory. Without clear constraints, it is hard to derive testable predictions. However, we can evaluate the specific model of linear orderings that Kemp and Tenenbaum (2008) provided. This model has two basic problems as a psychological proposal. First, given that the model can learn many different structural forms, it does not account for the empirical fact that linear orderings are special in the realm of animal cognition. As we have seen, a great variety of species can make comparative judgments based on linear orderings. By contrast, animals have considerably more difficulty learning circular orderings, or rings (von Fersen et al., 1991). The special status of linear orderings is a natural consequence for BARTlet and other models that base comparisons on magnitudes, or some similar unidimensional quantity. But within the Kemp and Tenenbaum framework, there is no apparent reason why rings should be any more difficult to learn than linear orders (though a prior could be arbitrarily imposed to favor either one).

A second basic problem is that the Kemp and Tenenbaum model of linear orders does not account for the ubiquitous distance effect. Their model creates explicit asymmetric relations between all possible pairs in an ordering (e.g., if elements A through E form a linear order, the learned structure would not only include links A > B, B > C, etc., but also B > D, B > E, etc.). The proposed inference processes (Kemp & Tenenbaum, 2009) imply that the strength of an inference concerning any two elements in a structure will be monotonic (in one direction or the other) with the length of the chain of links connecting the elements. But in the linear order model, the chain length is constant (one) for all pairs; hence the model predicts that (for example) a reasoner could evaluate B > C just as easily as B > D.

An ordering structure of this form was used to account for patterns of dominance behavior among members of a monkey troop (observed by Range & Noë, 2002; see Kemp & Tenenbaum, 2008, Fig. 4a, p. 10689). In fact, as we will discuss below, it is possible to explain monkeys' choices regarding whether or not to exhibit submissive behavior toward a conspecific without assuming that they form explicit comparative relations at all, far less a complete explicit representation of all pairwise relations. Thus, while the Kemp and Tenenbaum model of linear orders provides a useful tool for extracting the types of representations employed by (human) primatologists, it is problematic if interpreted as a psychological model of the mental representations that guide the choice behavior of primates.

5.6. Re-representation and the emergence of explicit relations

A great virtue of computational models is that they can bring clarity to important conceptual distinctions that might otherwise be blurred, or dismissed as a matter of semantics. A longstanding question in comparative psychology has been whether or not non-human animals (especially primates) "think", "reason", "use logic", or "understand relations" in fundamentally the same way as humans do. Various relational tasks have figured prominently as sources of evidence, including comparative judgment and transitive choice. As noted earlier, many species, from pigeons to primates, exhibit transitivity of choice (see Merritt & Terrace, 2011). Some have viewed such performance as tantamount to Piagetian transitive inference (e.g., if a 5-year old child is told that object B is bigger than object C, and object A is bigger than object B, then the child will likely be able to infer that A is bigger than C, despite knowing nothing about the features of the objects).

But in fact, transitivity of choice and Piagetian transitive inference involve completely different task demands, with little in common other than their misleadingly similar names (Halford, 1984; Markovits & Dumas, 1992). Transitivity of choice can be accomplished by using perceptually-based training data (ordered pairs and/or individual objects) to learn approximate quantities associated with individual items (e.g., magnitude codes, positional codes, values, or associative strengths). Examples of associative and statistical models that can accomplish learning of this type include the Rescorla-Wagner model (Rescorla & Wagner, 1972), Value Transfer Theory (von Fersen et al., 1991), RankSVM (Parikh & Grauman, 2011), and BARTlet. Although these models differ in many important ways, all provide mechanisms for performing relational judgments without explicit relations.

Accordingly, demonstrating success in basic comparative judgments tasks, or in transitivity of choice paradigms, cannot in principle provide evidence for the use of explicit comparative relations. Morgan's Canon can prudently be applied: "In no case may we interpret an action as the outcome of the exercise of a higher psychical faculty, if it can be interpreted as the outcome of the exercise of one which stands lower in the psychological scale" (Morgan, 1894, p. 53). If we replace the quaint

Victorian phrase "psychical faculty" with "relational complexity" or "representational rank" (Halford et al., 1998; Phillips, Halford, & Wilson, 1995), then Morgan's Canon continues to provide a valuable guide for comparative (and cognitive) psychology in the 21st century.

As Penn et al. (2008) argued based on a review of comparative studies, there is overwhelming evidence that many species of animals can make relational judgments based on perceptual information, yet no compelling evidence that any non-human primate is able to reason about relations. At the same time, it appears that the neural system supporting comparisons based on approximate magnitude in non-human primates operates in humans as well (Dehaene & Changeux, 1993). Apparently, humans have not lost the simpler mechanisms available to other animals for comparing magnitudes, but rather have exploited these mechanisms as a foundation for symbolic mathematical thinking (Opfer & Siegler, 2012). More generally, humans appear to have surpassed the intellectual capacity of any other species on earth by acquiring neural machinery that enables the re-representation of lower-level information in terms of explicit relational concepts.

As a small computational example of such re-representation, BARTlet becomes the prequel to BART, which uses one-place predicates such as *large* to bootstrap acquisition of explicit two-place relations such as *larger*. A system that is restricted to magnitude representations (lacking the ability to form explicit relational representations) inevitably "hits the wall" when faced with more complex symbolic tasks. A monkey (and BARTlet) can learn to choose the larger or the smaller of two objects. But a human (and BART) can also acquire an explicit representation of the relations *larger* and *smaller*, and go onto reason about them (e.g., noticing that *larger* is related to *smaller* in much the same way as *fiercer* is related to *meeker*; Lu et al., 2012).

Similarly, associative and statistical mechanisms that can support transitivity of choice prove completely inadequate when confronted with a Piagetian transitive inference task. The latter task requires a "one shot" inference based on integration of two binary premises in working memory, without repeated acquisition trials, and without support from perceptual cues or magnitude codes. Reliable success is not achieved by any species except humans, and not until preschool age (Andrews & Halford, 1998; Halford, 1984; Halford, 1993). Piagetian transitive inference is heavily dependent on a mature and intact human frontal cortex (Waltz et al., 1999). We have recently extended the BART model to enable it to use its learned representations to solve abstract transitive inference problems (Chen, Lu, & Holyoak, 2013). Perhaps surprisingly, explicit comparative relations are not required to make comparative judgments. However, they prove essential for any reasoner who aspires to think about what such judgments mean.

Acknowledgments

This research was supported by ONR Grant N000140810186. We thank Mark Steyvers for providing us with the code for the topic model, and Peter Gordon for providing us with a pre-processed copy of the Wikipedia corpus. In addition, we thank Robert Goldstone and two anonymous reviewers for helpful comments on an earlier draft. MATLAB code for the simulations reported here is available from the Web site of the UCLA Computational Vision and Learning Lab (http://cvl.psych.ucla.edu).

References

Andrews, G., & Halford, G. S. (1998). Children's ability to make transitive inferences: The importance of premise integration and structural complexity. *Cognitive Development, 13*(4), 479–513.

Audley, R. J., & Wallis, C. P. (1964). Response instructions and the speed of relative judgments: I. Some experiments on brightness discrimination. *British Journal of Psychology, 55*, 59–73.

Banks, W. P., Clark, H. H., & Lucy, P. (1975). The locus of the semantic congruity effect in comparative judgments. *Journal of Experimental Psychology: Human Perception and Performance, 104*, 35–47.

Banks, W. P., & Flora, J. (1977). Semantic and perceptual processes in symbolic comparisons. *Journal of Experimental Psychology: Human Perception and Performance, 3*, 278–290.

Banks, W. P., Fujii, M., & Kayra-Stuart, F. (1976). Semantic congruity effects in comparative judgments of magnitudes of digits. *Journal of Experimental Psychology: Human Perception and Performance, 2*, 435–447.

Banks, W. P., & White, H. (1982). Single ordering as a process limitation. *Journal of Verbal learning and Verbal Behavior, 21*, 39–54.

Banks, W. P., White, H., Sturgill, W., & Mermelstein, R. (1983). Semantic congruity and expectancy in symbolic judgments. *Journal of Experimental Psychology: Human Perception and Performance, 9*, 560–582.

Barner, D., & Snedeker, J. (2008). Compositionality and statistics in adjective acquisition: 4-year-olds interpret *tall* and *short* based on the size distributions of novel referents. *Child Development, 79*, 594–608.
Birnbaum, M. H., & Jou, J. W. (1990). A theory of comparative response times and "difference" judgments. *Cognitive Psychology, 22*, 184–210.
Bloom, P. (2000). *How children learn the meanings of words*. Cambridge, MA: MIT Press.
Bower, G. H. (1971). Adaptation-level coding of stimuli and serial position effects. In M. H. Appley (Ed.), *Adaptation-level theory* (pp. 175–201). New York: Academic Press.
Cantlon, J., & Brannon, E. M. (2005). Semantic congruity affects numerical judgments similarly in monkeys and humans. *Proceedings of the National Academy of Sciences, USA, 102*, 16507–16511.
Cantlon, J., Brannon, E. M., Carter, E., & Pelphrey, K. (2006). Functional imaging of numerical processing in adults and 4-yr-old children. *PLoS Biology, 4*(e125), 1–11.
Cantlon, J. F., Platt, M., & Brannon, E. M. (2009). Beyond the number domain. *Trends in Cognitive Sciences, 13*, 83–91.
Čech, C. G., & Shoben, E. J. (1985). Context effects in symbolic magnitude comparisons. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 11*, 299–315.
Čech, C., Shoben, E., & Love, M. (1990). Multiple congruity effects in judgments ofmagnitude. *Journal of Experimental Psychology Learning, Memory, and Cognition, 16*, 1142–1152.
Chen, D., Lu, H., & Holyoak, K. J. (2013). Generative inferences based on a discriminative Bayesian model of relation learning. In M. Knauf, M. Pauven, N. Sebanz, & I. Wachsmuth (Eds.), *Proceedings of the 35th annual conference of the Cognitive Science Society* (pp. 2018–2033). Austin, TX: Cognitive Science Society.
Clark, H. H. (1969). Linguistic processes in deductive reasoning. *Psychological Review, 1969*(76), 387–404.
Clark, H. H., Carpenter, P. A., & Just, M. A. (1973). On the meeting of semantics and perception. In W. G. Chase (Ed.), *Visual information processing* (pp. 311–381). New York: Academic Press.
Cromer, J. A., Roy, J. E., & Miller, E. K. (2010). Representation of multiple, independent categories in the primate prefrontal cortex. *Neuron, 66*, 796–807.
Davis, H. (1992). Transitive inference in rats (*Rattus norvegicus*). *Journal of Comparative Psychology, 106*, 342–349.
De Deyne, S., Verheyen, S., Ameel, E., Vanpaemel, W., Dry, M., Voorspoels, W., et al (2008). Exemplar by feature applicability matrices and other Dutch normative data for semantic concepts. *Behavior Research Methods, 40*(4), 1030–1048.
Dehaene, S. (1992). The varieties of numerical abilities. *Cognition, 44*, 1–42.
Dehaene, S., Bossini, S., & Giraux, P. (1993). The mental representation of parity and number magnitude. *Journal of Experimental Psychology: General, 122*, 371–396.
Dehaene, S., & Changeux, J.-P. (1993). Development of elementary numerical abilities: A neuronal model. *Journal of Cognitive Neuroscience, 5*, 390–407.
DeSoto, C. B. (1961). The predilection for single orderings. *Journal of Abnormal and Social Psychology, 62*, 16–23.
Diester, I., & Nieder, A. (2007). Semantic associations between signs and numerical categories in the prefrontal cortex. *PLoS Biology, 5*, e294.
Diester, I., & Nieder, Q. (2010). Numerical values leave a semantic imprint on associated signs in monkeys. *Journal of Cognitive Neuroscience, 22*, 174–183.
Dosher, B. A., & Lu, Z.-L. (2000). Noise exclusion in spatial attention. *Psychological Science, 11*, 139–146.
Doumas, L. A. A., Hummel, J. E., & Sandhofer, C. M. (2008). A theory of the discovery and predication of relational concepts. *Psychological Review, 115*, 1–43.
Freedman, D. J., Riesenhuber, M., Poggio, T., & Miller, E. K. (2001). Categorical representation of visual stimuli in the primate prefrontal cortex. *Science, 291*, 312–316.
Goldstone, R. L. (1994). The role of similarity in categorization: Providing a groundwork. *Cognition, 52*, 125–157.
Griffiths, T. L., Steyvers, M., & Tenenbaum, J. B. (2007). Topics in semantic representation. *Psychological Review, 114*, 211–244.
Halford, G. S. (1984). Can young children integrate premises in transitivity and serial order tasks? *Cognitive Psychology, 16*, 65–93.
Halford, G. S. (1993). *Children's understanding: The development of mental models*. Hillsdale, NJ: Erlbaum.
Halford, G. S., Wilson, W. H., & Phillips, S. (1998). Processing capacity defined by relational complexity: Implications for comparative, developmental and cognitive psychology. *Behavioral Brain Sciences, 21*(6), 803–831.
Halford, G. S., Wilson, W. H., & Phillips, S. (2010). Relational knowledge: The foundation of higher cognition. *Trends in Cognitive Sciences, 14*(11), 497–505.
Hoedemaker, R. S., & Gordon, P. C. (2013). Embodied language comprehension: Encoding-based and goal-driven processes. *Journal of Experimental Psychology: General*. http://dx.doi.org/10.1037/a0032348.
Holyoak, K. J. (1977). The form of analog size information in memory. *Cognitive Psychology, 9*, 31–51.
Holyoak, K. J. (1978). Comparative judgments with numerical reference points. *Cognitive Psychology, 10*, 203–243.
Holyoak, K. J., Dumais, S. T., & Moyer, R. S. (1979). Semantic association effects in a mental comparison task. *Memory & Cognition, 7*, 303–313.
Holyoak, K. J., & Mah, W. A. (1981). Semantic congruity in symbolic comparisons: Evidence against an expectancy hypothesis. *Memory & Cognition, 9*, 197–204.
Holyoak, K. J., & Mah, W. A. (1982). Cognitive reference points in judgments of symbolic magnitude. *Cognitive Psychology, 14*, 328–352.
Holyoak, K. J., & Patterson, K. K. (1981). A positional discriminability model of linear order judgments. *Journal of Experimental Psychology: Human Perception and Performance, 7*, 1283–1302.
Holyoak, K. J., & Walker, J. H. (1976). Subjective magnitude information in semantic orderings. *Journal of Verbal Learning and Verbal Behavior, 15*, 287–299.
Howard, R. (1983). The semantic congruity effect: Some tests of the expectancy hypothesis. *Acta Psychologica, 53*, 205–216.
Jamieson, D. G., & Petrusic, W. (1975). Relational judgments with remembered stimuli. *Perception & Psychophysics, 18*, 373–378.
Jones, S. M., Cantlon, J. F., Merritt, D. J., & Brannon, E. M. (2010). Context affects the numerical semantic congruity effect in rhesus monkeys (*Macaca mulatta*). *Behavioral Processes, 83*, 191–196.

Kemp, C., & Tenenbaum, J. B. (2008). The discovery of structural form. *Proceedings of the National Academy of Sciences, USA, 105*(31), 10687–10692.
Kemp, C., & Tenenbaum, J. B. (2009). Structured statistical models of inductive reasoning. *Psychological Review, 116*(1), 20–58.
Lawrence, D. H., & DeRivera, J. (1954). Evidence for relational transposition. *Journal of Comparative and Physiological Psychology, 47*, 465–471.
Link, S. W. (1990). Modeling imageless thought: The relative judgment theory of numerical comparisons. *Journal of Mathematical Psychology, 34*, 2–41.
Lu, H., Chen, D., & Holyoak, K. J. (2012). Bayesian analogy with relational transformations. *Psychological Review, 119*, 617–648.
Luce, R. D., Green, D. M., & Weber, D. L. (1976). Attention bands in absolute identification. *Perception & Psychophysics, 20*, 49–54.
Markovits, H., & Dumas, C. (1992). Can pigeons really make transitive inferences? *Journal of Experimental Psychology: Animal Behavior Processes, 18*, 311–312.
Marks, D. F. (1972). Relative judgment: A phenomenon and a theory. *Perception & Psychophysics, 11*, 156–160.
Marr, D. (1982). *Vision: A computational investigation into the human representation and processing of visual information*. New York: W. H. Freeman.
Marschark, M., & Paivio, A. (1979). Semantic congruity and lexical marking in symbolic comparisons: An expectancy hypothesis. *Memory & Cognition, 7*, 175–184.
McGonigle, B. O., & Chalmers, M. (1977). Are monkeys logical? *Nature, 267*, 694–696.
Merritt, D. J., & Terrace, H. S. (2011). Mechanisms of inferential order judgments in humans (*Homo sapiens*) and rhesus monkeys (*Macaca mulatta*). *Journal of Comparative Psychology, 125*, 227–238.
Miller, G. A. (1956). The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological Review, 63*, 81–97.
Morgan, C. L. (1894). *An introduction to comparative psychology*. London: Walter Scott.
Moyer, R. S. (1973). Comparing objects in memory: Evidence suggesting an internal psychophysics. *Perception & Psychophysics, 13*, 180–184.
Moyer, R. S., & Bayer, R. H. (1976). Mental comparison and the symbolic distance effect. *Cognitive Psychology, 8*, 228–246.
Moyer, R. S., & Dumais, S. T. (1978). Mental comparison. In G. H. Bower (Ed.). *The psychology of learning and motivation* (Vol. 12, pp. 117–155). New York: Academic Press.
Moyer, R. S., & Landauer, T. K. (1967). Time required for judgments of numerical inequality. *Nature, 215*(1519–1), 520.
Murdock, B. B. Jr., (1960). The distinctiveness of stimuli. *Psychological Review, 67*, 16–31.
Nagpal, A., & Krishnamurthy, P. (2008). Attribute conflict in consumer decision making: The role of task compatibility. *Journal of Consumer Research, 34*, 696–705.
Nieder, A., & Miller, E. K. (2003). Coding of cognitive magnitude: Compressed scaling of numerical information in the primate prefrontal cortex. *Neuron, 37*, 149–157.
Nosofsky, R. M. (1983). Shifts of attention in the identification and discrimination of intensity. *Perception & Psychophysics, 33*, 103–112.
Opfer, J. E., & Siegler, R. S. (2012). Development of quantitative thinking. In K. K. Holyoak & R. G. Morrison (Eds.), *Oxford handbook of thinking and reasoning* (pp. 585–605). New York: Oxford University Press.
Parikh, D., & Grauman, K. (2011). Relative attributes. In D. M. Metaxas, L. Quan, & L. J. Van (Eds.), *Proceedings of the IEEE international conference on computer vision* (pp. 503–510). Barcelona, Spain: IEEE.
Partee, B. (1995). Lexical semantics and compositionality. In D. Osherson (General Ed.), & L. Gleitman, & M. Liberman (Eds.), *Invitation to cognitive science. Part I: Language* (2nd ed., pp. 311–360). Cambridge, MA: MIT Press.
Penn, D. C., Holyoak, K. J., & Povinelli, D. J. (2008). Darwin's mistake: Explaining the discontinuity between human and nonhuman minds. *Behavioral and Brain Sciences, 31*, 109–178.
Petrusic, W. M. (1992). Semantic congruity effects and theories of the comparison process. *Journal of Experimental Psychology: Human Perception and Performance, 18*, 962–986.
Petrusic, W. M., & Baranski, J. V. (1989). Semantic congruity effects in perceptual comparisons. *Perception & Psychophysics, 45*, 439–452.
Phillips, S., Halford, G. S., & Wilson, W. H. (1995). The processing of associations versus the processing of relations and symbols: A systematic comparison. In J. D. Moore & J. F. Lehman (Eds.), *Proceedings of the seventeenth annual conference of the Cognitive Science Society* (pp. 688–691). Mahwah, NJ: Erlbaum.
Piazza, M., Izard, V., Pinel, P., Bihan, D., & Dehaene, S. (2004). Tuning curves for approximate numerosity in the human intraparietal sulcus. *Neuron, 44*, 547–555.
Piazza, M., Mechelli, A., Price, C. J., & Butterworth, B. (2006). Exact and approximate judgements of visual and auditory numerosity: An fMRI study. *Brain Research, 1106*, 177–188.
Piazza, M., Pinel, P., Le Bihan, D., & Dehaene, S. (2007). A magnitude code common to numerosities and number symbols in human intraparietal cortex. *Neuron, 53*, 293–305.
Pinel, P., Piazza, M., Bihan, D. L., & Dehaene, S. (2004). Distributed and overlapping cerebral representations of number, size, and luminance during comparative judgments. *Neuron, 41*, 1–20.
Potts, G. R. (1974). Storing and retrieving information about ordered relationships. *Journal of Experimental Psychology, 103*, 431–439.
Rahnev, D., Maniscalco, B., Graves, T., Huang, E., de Lange, F. P., & Lau, H. (2011). Attention induces conservative subjective biases in visual perception. *Nature Neuroscience, 14*, 1513–1515.
Range, F., & Noë, R. (2002). Familiarity and dominance relations among female sooty mangabeys in the Taï National Park. *American Journal of Primatology, 56*, 137–153.
Ratcliff, R. (1978). A theory of memory retrieval. *Psychological Review, 85*, 59–108.
Ratcliff, R., Cherian, A., & Segraves, M. (2003). A comparison of macaque behavior and superior colliculus neuronal activity to predictions from models of simple two-choice decisions. *Journal of Neurophysiology, 90*, 1392–1407.
Ratcliff, R., & McKoon, G. (2010). The diffusion decision model: Theory and data for two-choice decision tasks. *Neural Computation, 20*, 873–922.

Ratcliff, R., Van Zandt, T., & McKoon, G. (1999). Connectionist and diffusion models of reaction time. *Psychological Review, 106*, 261–300.
Rescorla, R. A., & Wagner, A. R. (1972). A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. In A. H. Black & W. F. Prokasy (Eds.), *Classical conditioning II: Current research and theory* (pp. 64–99). New York: Appleton-Century-Crofts.
Reynolds, J. H., & Chelazzi, R. (2004). Attentional modulation of visual processing. *Annual Review of Neuroscience, 27*, 611–647.
Ryalls, B. O., Winslow, E., & Smith, L. B. (1998). A semantic congruity effect in children's acquisition of *high* and *low*. *Journal of Memory and Language, 39*, 543–557.
Rylass, B. O., & Smith, L. B. (2000). Adults' acquisition of novel dimension words: Creating a semantic congruity effect. *Journal of General Psychology, 127*, 279–326.
Shafto, P., Kemp, C., Mansinghka, V., & Tenenbaum, J. B. (2011). A probabilistic model of cross-categorization. *Cognition, 120*, 1–25.
Shaki, S., Leth-Steensen, C., & Petrusic, W. W. (2006). Effects of instruction presentation mode in comparative judgment. *Memory & Cognition, 34*, 196–206.
Shaki, S., Petrusic, W. M., & Leth-Steensen, C. (2012). SNARC effects with numerical and non-numerical symbolic comparative judgments: Instructional and cultural dependencies. *Journal of Experimental Psychology: Human Perception and Performance, 38*, 515–530.
Shepard, R. N., Kilpatric, D. W., & Cunningham, J. P. (1975). The internal representation of numbers. *Cognitive Psychology, 7*, 82–138.
Shoben, E. J., Sailor, K. M., & Wang, M. (1989). The role of expectancy in comparative judgments. *Memory & Cognition, 17*, 18–26.
Smith, L. B., Gasser, M., & Sandhofer, C. M. (1997). Learning to talk about the properties of objects: A network model of the development of dimensions. In R. L. Goldstone, D. L. Medin, & P. G. Schyns (Eds.), *Advances in the psychology of learning and motivation. Perceptual learning* (Vol. 36, pp. 219–255). San Diego, CA: Academic Press.
Trabasso, T., & Riley, C. A. (1975). The construction and use of representations involving linear order. In R. L. Solso (Ed.), *Information processing and cognition: The Loyola symposium* (pp. 381–410). Hillsdale, NJ: Erlbaum.
Tribushinina, E. (2009). Reference points in linguistic construal: Scalar adjectives revisited. *Studia Linguistica, 63*, 233–260.
Tversky, A., & Kahneman, D. (1981). The framing of decisions and the psychology of choice. *Science, 211*, 453–458.
von Fersen, L., Wynne, C. D. L., Delius, J. D., & Staddon, J. E. (1991). Transitive inference in pigeons. *Journal of Experimental Psychology: Animal Behavior Processes, 17*, 334–341.
Wallis, C. P., & Audley, R. J. (1964). Response instructions and the speed of relative judgments: II. Pitch discrimination. *British Journal of Psychology, 55*, 121–132.
Waltz, J. A., Knowlton, B. J., Holyoak, K. J., Boone, K. B., Mishkin, F. S., de Menezes Santos, M., et al (1999). A system for relational reasoning in human prefrontal cortex. *Psychological Science, 10*(2), 119–125.
Wickens, T. D. (2002). *Elementary signal detection theory*. New York: Oxford University Press.
Wong, K.-F., & Wang, X.-J. (2006). A recurrent network mechanism of time integration in perceptual decisions. *Journal of Neuroscience, 26*, 1314–1328.
Woocher, F. D., Glass, A. L., & Holyoak, K. J. (1978). Positional discriminability in linear orderings. *Memory & Cognition, 6*, 165–173.
Wynne, C. D. L. (1995). Reinforcement accounts for transitive inference performance. *Animal Learning & Behavior, 23*, 207–217.