Optimal inference of sameness

Ronald van den Berg[a], Michael Vogel[b], Krešimir Josić[b,c], and Wei Ji Ma[a,1]

[a]Department of Neuroscience, Baylor College of Medicine, Houston, TX 77030; and Departments of [b]Mathematics and [c]Biology and Biochemistry, University of Houston, Houston, TX 77204

Edited by Wilson S. Geisler, University of Texas, Austin, TX, and approved December 8, 2011 (received for review June 5, 2011)

Deciding whether a set of objects are the same or different is a cornerstone of perception and cognition. Surprisingly, no principled quantitative model of sameness judgment exists. We tested whether human sameness judgment under sensory noise can be modeled as a form of probabilistically optimal inference. An optimal observer would compare the reliability-weighted variance of the sensory measurements with a set size-dependent criterion. We conducted two experiments, in which we varied set size and individual stimulus reliabilities. We found that the optimal-observer model accurately describes human behavior, outperforms plausible alternatives in a rigorous model comparison, and accounts for three key findings in the animal cognition literature. Our results provide a normative footing for the study of sameness judgment and indicate that the notion of perception as near-optimal inference extends to abstract relations.

Bayesian inference | ideal observer | decision making | vision

According to William James, the "sense of sameness is the very keel and backbone of our thinking" (1). Judging whether a set of stimuli are all the same (in one or multiple features) indeed seems to be a fundamental component of perception and cognition. When segmenting a visual scene, we make use of the fact that features such as orientation and color tend to be the same within an object and different between objects (2). Learning to categorize objects requires evaluating the aspects in which they are the same or different. Sameness judgment is also thought to play a role in the development of the abstract notion of equivalence, fundamental to learning mathematics (3). The central role of sameness judgment in perception and cognition is reflected in its prevalence in many psychophysical tasks, including matching to sample (4), change detection (5, 6), oddity search (7, 8), and causal inference (9).

The ability to judge sameness seems to have a long evolutionary history. In addition to humans, honey bees (10), pigeons (11), parrots (12), dolphins (13), Old World monkeys (5), New World monkeys (14, 15), and apes (6, 7) can all learn to report whether pairs of objects are the same or different. This ability extends to larger groups of objects in both humans and nonhuman animals (16–21). It has been suggested that human capacity to recognize sameness at abstract levels is closely linked to the evolution of prefrontal cortex (22).

In animal cognition experiments on sameness judgment, several intriguing trends have not yet been fully explained. First, when pigeons are trained to discriminate arrays of identical objects from arrays of different objects, their probability of reporting "different" is found to gradually increase with the amount of variability in the array (23, 24). For example, when an array contained four subsets of four icons each, with icons being identical within a subset, subjects reported "different" more frequently than when two subsets of eight icons were shown. Second, sameness judgment becomes easier with increasing number of stimuli (set size) for both pigeons and baboons (16, 17). Third, when stimuli are blurred, pigeons more frequently respond "same" on "different" trials, but performance is more or less unchanged on "same" trials (20).

Despite the ecological importance of sameness judgment and the abundance of experimental data, the computations underlying sameness judgment are not well understood. According to one proposal, organisms assess sameness of a set of simultaneously presented items by estimating the entropy of the set (24), but this turned out to be inadequate as a general model (20). The more recent "finding differences" model (20) posits that sameness judgment is based on local differences between items. Although this model is a step forward in terms of fitting behavioral data, it is descriptive and lacks a normative basis: it does not explain *why* subjects base their sameness judgments on an estimate of stimulus variability, and it postulates rather than derives the observer's mapping from stimuli to response probabilities.

Here we propose that observers attempt to maximize performance in their judgments of sameness in the presence of sensory noise in the measurements. This simple principle leads to a precise mathematical model in which no ad hoc assumptions are needed—the optimal-observer model. As we will see, this model predicts that observers judge sameness by using the reliability-weighted variance of the sensory measurements as their decision variable. We find that this model provides an accurate quantitative description of human data, outperforms alternative models, and accounts for the above-mentioned animal cognition findings.

Experiment 1: Varying Set Size

In experiment 1, observers ($n = 8$, with 2,700 trials each) were asked to judge whether the orientations of a set of ellipses simultaneously presented for 67 ms were the same or different (Fig. 1 *A* and *B*). Set size was 2, 4, or 8 in separate blocks. "Same" and "different" trials both occurred with probability 0.5. On each trial, an orientation μ was drawn from a uniform distribution over all possible orientations. On a "same" trial, the orientations of the individual ellipses were all equal to μ. On a "different" trial, these orientations were drawn from a Gaussian distribution with mean μ and SD $\sigma_s = 10°$.

The proportion of "different" responses is shown as a function of set size in Fig. 1*C*. Performance improves on both "same" and "different" trials as more stimuli are presented. The effect of set size was highly significant for both trial types [repeated-measures ANOVA; $F(2,14) = 15.4$, $P < 0.001$, and $F(2,14) = 120$, $P < 0.001$, respectively]. Even though on "different" trials, orientations are always drawn using the same SD of $\sigma_s = 10°$, the SD of the actually presented orientations varies widely from trial to trial. Therefore, a more detailed representation of the data is provided by the proportion of "different" responses as a function of the sample SD of the presented stimuli (Fig. 1*D*; normalized by $N - 1$). We observe a clear effect of sample SD.

Optimal-Observer Model. The statistical structure of the task (generative model) is illustrated in Fig. 2*A*. We denote the orientation of the stimulus at the ith location by s_i, and model its

Author contributions: R.v.d.B., K.J., and W.J.M. designed research; R.v.d.B., M.V., and W.J.M. performed research; R.v.d.B. and M.V. analyzed data; and R.v.d.B., K.J., and W.J.M. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

[1]To whom correspondence should be addressed. E-mail: wjma@bcm.edu.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1108790109/-/DCSupplemental.

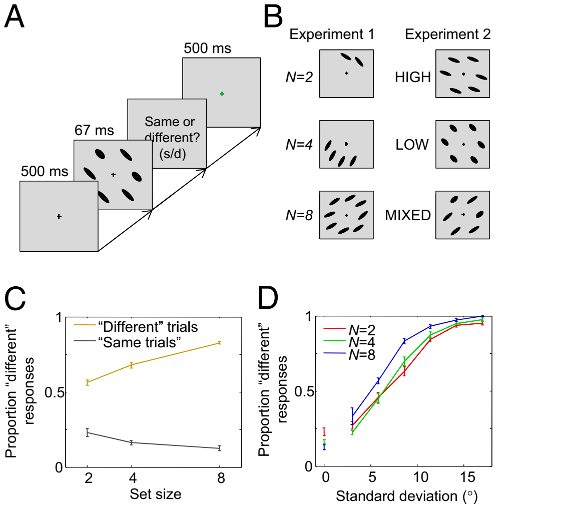

Fig. 1. Experimental procedure and psychometric curves. Error bars indicate SEM over subjects. (*A*) Subjects reported through a key press whether the orientations of the ellipses were identical. (*B*) Experimental conditions. In experiment 1, set size was 2, 4, or 8. In experiment 2, set size was 6 and stimuli all had high reliability (HIGH), all low reliability (LOW), or were mixed (MIXED). (*C*) Proportion of "different" responses in experiment 1 as a function of set size. (*D*) Proportion of "different" responses in experiment 1 as a function of the SD of the presented orientations. All "same" trials have an SD of 0.

measurement by the observer, denoted x_i, as a Gaussian random variable with mean s_i and SD $\sigma\sigma$.) Judging whether the orientations of a set of N stimuli, $\mathbf{s} = \{s_1,\ldots,s_N\}$, are the same or different, amounts to inferring sameness ($C = 1$ for same, $C = -1$ for different) from $\mathbf{x} = \{x_1,\ldots,x_N\}$. A probabilistically optimal observer computes, on each trial, the posterior probability that the stimuli have equal orientation, $p(C = 1|\mathbf{x})$, and responds "same" when this probability exceeds 0.5. This probability is obtained from the likelihood function over C, $p(\mathbf{x}|C)$, that is, the probability of the measurements, $\mathbf{x}$, if the true state of the world is C. Because the true orientations, $\mathbf{s}$, and their mean, μ, are unknown to the observer, the likelihood is computed by averaging over all possible values of these quantities, a computation known as *marginalization*:

$$p(\mathbf{x}|C) = \iint p(\mathbf{x}|\mathbf{s})\,p(\mathbf{s}|C,\mu)\,p(\mu)\,d\mathbf{s}\,d\mu.$$

The optimal observer responds "same" when the posterior probability $p(C = 1|\mathbf{x})$ is greater than $p(C = -1|\mathbf{x})$, or, equivalently, if $\log\frac{p(C=1|\mathbf{x})}{p(C=-1|\mathbf{x})} > 0$. This condition translates into a condition on the variance of the measurements (*SI Appendix*),

$$\text{Var}\,\mathbf{x} < \frac{1}{N(w-\tilde{w})}\left[(N-1)\log\frac{w}{\tilde{w}} + 2\log\frac{p_{\text{same}}}{1-p_{\text{same}}}\right], \qquad [1]$$

where $p_{\text{same}} = p(C = 1)$ is the observer's prior probability of sameness, $w = 1/\sigma^2$, and $\tilde{w} = 1/(\sigma^2 + \sigma_s^2)$. Eq. **1** states that the optimal strategy is to report "same" when the variance of the measurements, Var $\mathbf{x}$, is smaller than a decision criterion that depends on set size, reliability w, and prior probability of sameness. Larger Var $\mathbf{x}$ implies a higher probability that the observed differences are not only due to internal noise but also due to differences among the stimuli. The optimal decision boundary is formed by the points $\mathbf{x}$ for which Eq. **1** is an equality (Fig. 2*B*). In many binary classification tasks, the optimal decision boundary is a plane in a high-dimensional space. By contrast, the left-hand side of Eq. **1** is a quadratic form in $\mathbf{x}$, and the corresponding decision boundary is a cylinder. This is an indication of the complexity of this inference problem.

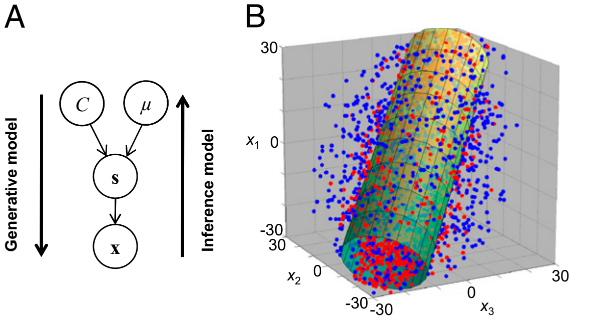

Fig. 2. Statistical structure of the task and a geometrical interpretation of the optimal decision rule. (*A*) Each node represents a random variable, each arrow a conditional probability distribution. This diagram specifies the distribution of measurements, **x**. The optimal observer "inverts" the generative model and computes the probability of *C* given **x**. (*B*) Geometrical interpretation of the optimal decision rule at *N* = 3. The axes represent the observed stimulus orientations, **x**. Each dot represents the set of measurements on a single trial. On "same" trials (red), the dots lie—on average—closer to the diagonal than on "different" trials (blue). The optimal strategy is to respond "same" when **x** lies within the green cylinder, whose axis is the diagonal.

We let the noise level at each set size, σ, be a free parameter. Three further parameters define multiple variants of the optimal model. For the prior over sameness, the observer could be using the correct value ($p_{\text{same}} = 0.5$) or a different value. To account for the latter possibility, we consider a model variant in which p_{same} is a free parameter. Similarly, for the theoretical SD of "different" stimuli, the observer could be using the correct value, $\sigma_s = 10°$, or a different value. Finally, blinking, lapses in attention, and other factors may cause subjects to guess randomly on some trials. Therefore, there might or might not be a nonzero lapse rate. In total, we therefore consider 2 × 2 × 2 = 8 variants with anywhere from three to six free parameters (*SI Appendix*, Table S1).

Alternative Models. We compare the optimal model against the following suboptimal models. In the *single-criterion (SC) model*, the observer uses the decision rule Var $\mathbf{x} < k$, where k is independent of set size, reliability, and prior (it is a free parameter in the model). This model has four free parameters. In the *blockwise-criterion (BC) model*, the observer uses one criterion per block type. Because blocks only differed by set size, the decision rule is Var $\mathbf{x} < k_N$, where k_N is the decision criterion at set size N. At the cost of having more free parameters (3 σs and 3 ks), the BC model has more freedom in fitting the decision criteria than the optimal model, in which the decision criteria at different set sizes are linked through Eq. **1**. We will compare the fitted BC and optimal criteria. In the *maximum-absolute-difference (MAD) model* (21), the decision rule is to respond "same" if the largest unsigned difference between any two measurements is smaller than a criterion k (i.e., $\max\limits_{i,j}|x_i - x_j| < k$). Again, this criterion may or may not differ by block type, leading to two models: MAD with a single criterion (MAD-SC, four free parameters) and MAD with a blockwise criterion (MAD-BC, six free parameters). Finally, each of the alternative models may or may not have a lapse rate, which means that each of them comes in two variants (*SI Appendix*, Table S1).

Results. For each model, we computed the probability of reporting "different" on each trial given the stimuli on that trial. From this, we computed the probability of all of a subject's responses given a model and its parameters (*SI Appendix*). We fitted parameters by maximizing this parameter likelihood. Model fits to subject data are shown in Fig. 3*A*. For each subject and each model, we used the maximum-likelihood fit from the most likely model variant. Comparing the RMS errors, the optimal model (0.032) fits approximately as well as the BC (0.029) and MAD-BC (0.029) models and much better than the SC (0.087) and the MAD-SC models (0.136).

We computed model likelihoods using Bayesian model comparison, a method that uses the stimulus–response pairs on individual trials and automatically takes into account differences between models in numbers of free parameters (*SI Appendix*). For each model, we averaged the likelihoods of all its variants. The differences in log likelihood between the optimal model and the SC, BC, MAD-SC, and MAD-BC models were 25.7 ± 5.3, −0.70 ± 2.3, 63.5 ± 8.3, and 5.06 ± 2.5, respectively (Fig. 3*B*). Hence, the optimal model and the BC models describe the data better than the SC models. However, these data cannot distinguish between the optimal and the BC models.

The optimal model is a special case of the BC model: in the former model, the decision criterion has a prescribed dependence on set size, whereas in the latter, it is a free parameter at each set size. Hence, if subjects followed the optimal strategy, then we should find that the BC model, despite its greater freedom in setting the criterion, closely approximates the optimal model when fitted to the subject data. This is exactly what we found. The decision criteria from the best BC model variant are nearly identical to those predicted by the best optimal-model variant (Fig. 3*C*). (Different variants of the optimal model fit best for different subjects; *SI Appendix*, Table S2) Paired t tests did not reveal a significant difference at any of the three set sizes ($P = 0.09$, $P = 0.12$, and $P = 0.58$ at $N = 2$, 4, and 8, respectively). Similarly, the fitted noise levels in the BC and the optimal model are also nearly identical (Fig. 3*C*), with no significant difference in any of the set sizes ($P = 0.72$, $P = 0.49$, and $P = 0.67$ at $N = 2$, 4, and 8, respectively). The noise level exhibits a weak dependence on set size, with a power of 0.16 ± 0.03 in a power-law fit. We conclude that the decision criteria used by the human subjects were close to optimal. To assess the generality of these results, we repeated the experiment by varying a different stimulus feature, namely color. The pattern of results is similar (*SI Appendix*).

Experiment 2: Varying Stimulus Reliabilities

To further distinguish the models, specifically the optimal model from both BC models, we tested a prediction unique to the optimal model, namely that the observer weights the measurements within a single display by their respective reliabilities (25, 26). This can be tested by experimentally varying stimulus reliability across stimuli in the same display. In experiment 2, we manipulated stimulus reliability through the eccentricity (elongation) of the ellipse (27) (Fig. 1*B*; *Materials and Methods*). Two eccentricity values were used, which we call low and high. Three reliability conditions were presented in separate blocks: LOW, HIGH, and MIXED. In the LOW and HIGH conditions, all ellipses had low and high eccentricity, respectively. In the MIXED condition, the eccentricity of each ellipse on each trial was set to the high or the low value independently and with equal probability. The experimental procedure was identical to experiment 1. Ten observers each completed 2,700 trials.

Optimal Model. In the optimal decision rule in experiment 1, Eq. **1**, the observer compared the variance of the measurements with a criterion. Here, each term entering in the variance is weighted by the reliability of the respective measurement. Evidence from an elongated ellipse is weighted more heavily than that of a more circular one. The resulting "reliability-weighted variance" is compared with a criterion that depends on the stimulus reliabilities and the prior (*SI Appendix*). In the N-dimensional space of measurements, the optimal decision boundary is an elliptic cylinder, with the size and shape of the ellipsoidal cross section depending on the specific set of reliabilities on a given trial. In the MIXED condition, these reliabilities vary from trial to trial, and the decision boundary reshapes accordingly.

Alternative Models. We consider the same four suboptimal models as in experiment 1, except that in the BC and MAD-BC models, the decision criterion now varies not by set size but by reliability condition (LOW, HIGH, or MIXED). Thus, the respective decision rules are $\mathrm{Var}\,\mathbf{x} < k_{\mathrm{block}}$ and $\max_{i,j} |x_i - x_j| < k_{\mathrm{block}}$, where k_{block} is the decision criterion for a given block type (this introduces three free parameters). None of the alternative observers take into account item-to-item reliability.

Results. In each reliability condition, the optimal model accounts well for the proportion of "different" responses as a function of sample SD (Fig. 4*A*). It fits the data better than any of the alternative models (RMS errors: optimal: 0.024; SC: 0.074; BC: 0.062; MAD-SC: 0.075; MAD-BC: 0.035). The noise parameters corresponding to high- and low-eccentricity ellipses were estimated to be 6.1° ± 0.2° and 10.7° ± 0.4°, respectively.

To distinguish the models further, we grouped trials by the number of high-reliability stimuli, denoted N_{high} (0 in LOW, 0–6 in MIXED, 6 in HIGH). We found that the proportion of "different" reports increased with N_{high} on "different" trials, whereas it decreased on "same" trials (Fig. 4*B*). These effects

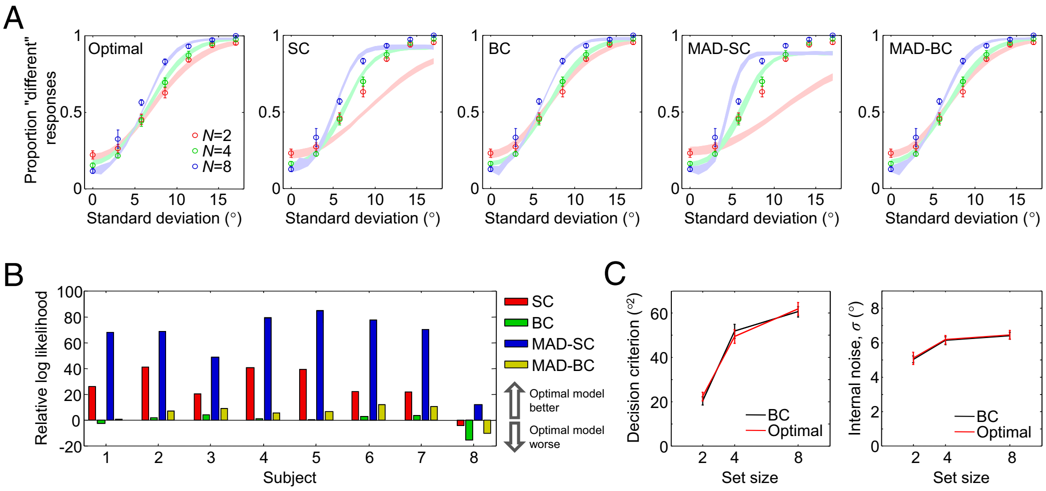

Fig. 3. Comparison of models in experiment 1. Circles and error bars represent mean and SEM of subject data. Shaded areas represent SEM of model fits. (*A*) Proportion "different" responses as a function of sample SD for optimal and suboptimal models. (*B*) Bayesian model comparison. Each bar represent the log likelihood of the optimal model minus that of a suboptimal model. (*C*) The decision criteria (*Left*) and the internal noise levels (*Right*) for the best-fitting BC model are nearly identical to those of the best-fitting optimal-observer model. This suggests that the human criteria are close to optimal.

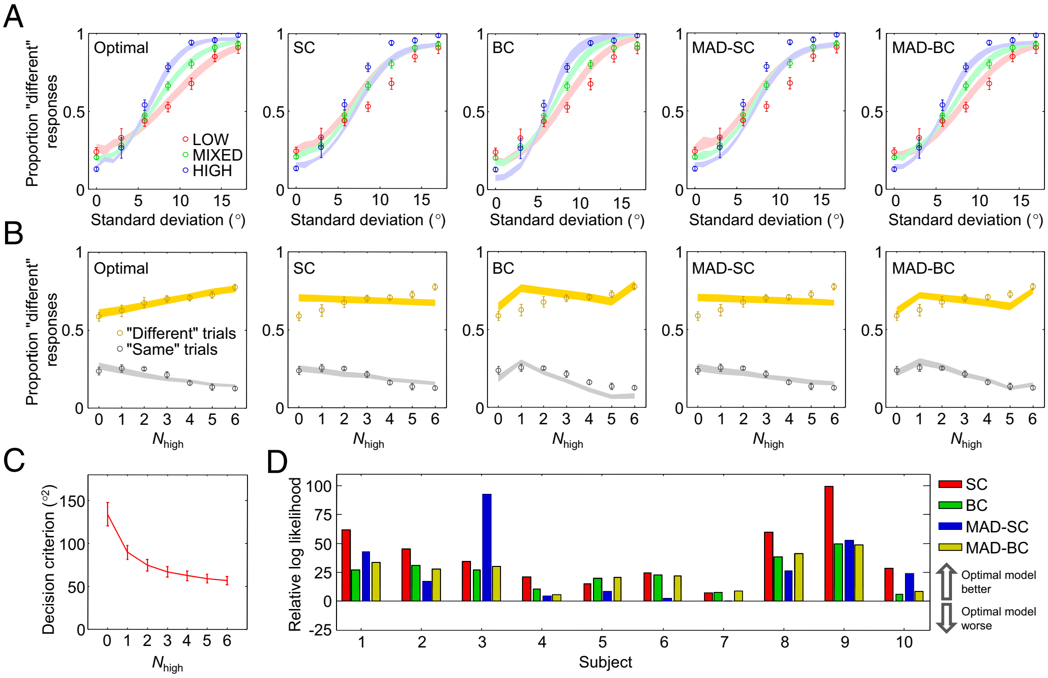

Fig. 4. Comparison of models in experiment 2. Circles and error bars represent mean and SEM of subject data. Shaded areas represent SEM of model fits. (*A*) Proportion "different" responses as a function of sample SD for optimal and suboptimal models. (*B*) Proportion of "different" responses as a function of the number of high-reliability stimuli in a display (N_{high}). In both BC models, LOW ($N_{high} = 0$) and HIGH ($N_{high} = 6$) were fitted separately. (*C*) Decision criterion in the optimal model as a function of N_{high}. (*D*) Bayesian model comparison. Each bar represents the log likelihood of the optimal model minus that of a suboptimal model.

were significant for both trial types [repeated-measures ANOVA: $F(6,54) = 15.8$, $P < 0.001$, and $F(6,54) = 19.8$, $P < 0.001$, respectively]. The optimal model provides the best fit to these data (RMS errors: optimal: 0.020; SC: 0.052; BC: 0.058; MAD-SC: 0.052; MAD-BC: 0.039) (Fig. 4*B*). On "different" trials in the MIXED condition, the optimal model makes a qualitatively different prediction from all other models, namely that the proportion of "different" responses increases with N_{high}. This can be understood from the distributions of the decision variable (*SI Appendix*, Fig. S2). As these distributions change with N_{high}, the optimal observer adjusts the decision criterion accordingly (Fig. 4*C* and *SI Appendix*, Fig. S2), but the alternative observers keep using the same criterion. The proportion of "different" responses on "different" trials increases with N_{high} in the optimal model because the distributions become better separable, and decreases with N_{high} in the alternative models because the distributions shift to lower values.

Bayesian model comparison shows that the optimal model outperforms the SC, BC, MAD-SC, and MAD-BC models by 39.6 ± 8.7, 23.9 ± 4.4, 41.2 ± 8.6, and 24.7 ± 4.6 log likelihood points, respectively. Moreover, the optimal model best describes the behavior of each individual subject (Fig. 4*D*). Different variants of the optimal model fit best for different subjects (*SI Appendix*, Table S2).

All suboptimal models considered so far used a decision criterion that was constant within a block. The optimal model was the only model in which the decision criterion changed from trial to trial. A very strong test of the optimal model would be to compare it against a suboptimal model in which the decision criterion is allowed to vary from trial to trial. We implemented such a model by assuming that the observer averages, on each trial, the reliabilities of the stimuli on that trial and uses this average in the decision rule. This rule is equivalent to Var $\mathbf{x} < k$, where k depends in a specific way on N_{high}. Unlike in the optimal model, the measurements, x_i, are not individually weighted by their respective reliabilities. We found that this suboptimal model is less likely than the Bayesian model by 27.0 ± 9.1 log likelihood points. Together, our results constitute decisive evidence in favor of the optimal model and indicate that humans weight measurements by reliability when judging sameness.

Reexamination of Sameness Judgments by Animals

We examined three findings from the animal cognition literature on sameness judgment from the perspective of the optimal-observer model (details in *SI Appendix*). First, Young et al. (24), using picture stimuli, reported that the probability that pigeons respond "different" strongly correlates with the entropy of the stimulus set (Fig. 5*A*, *Upper*). We simulated the optimal observer on an equivalent task with orientation stimuli and found that the responses show the same strong correlation (Fig. 5*A*, *Lower*). In the optimal-observer model, the intuition behind this correlation is that the probability of responding "different" depends on the variance of the stimulus set, and entropy is correlated with variance. Thus, instead of detecting entropy per se (24), pigeons might be using a decision rule similar to the optimal observer when judging sameness.

Next we tested whether the optimal-observer model can account for the effects of set size on animals' reports of sameness. Wasserman and colleagues found that proportion correct increased with set size on both "same" and "different" trials (Fig. 5*B*, *Upper*) (16, 17). The optimal-observer model reproduces these effects (Fig. 5*B*, *Lower*). The slightly steeper slopes seen in the data could be due to the fact that in picture stimuli more than one feature can be used to determine sameness, leading to a greater "effective" set size. The effects of set size can be understood by recognizing that optimal observers use the variance of the measurements, Var $\mathbf{x}$, to discriminate "same" from "different" (Eq. **1**). At larger set sizes, $\mathbf{x}$ contains more elements and therefore Var $\mathbf{x}$ will vary less from trial to trial. This, in turn, will lead to less overlap between the distributions of Var $\mathbf{x}$ on

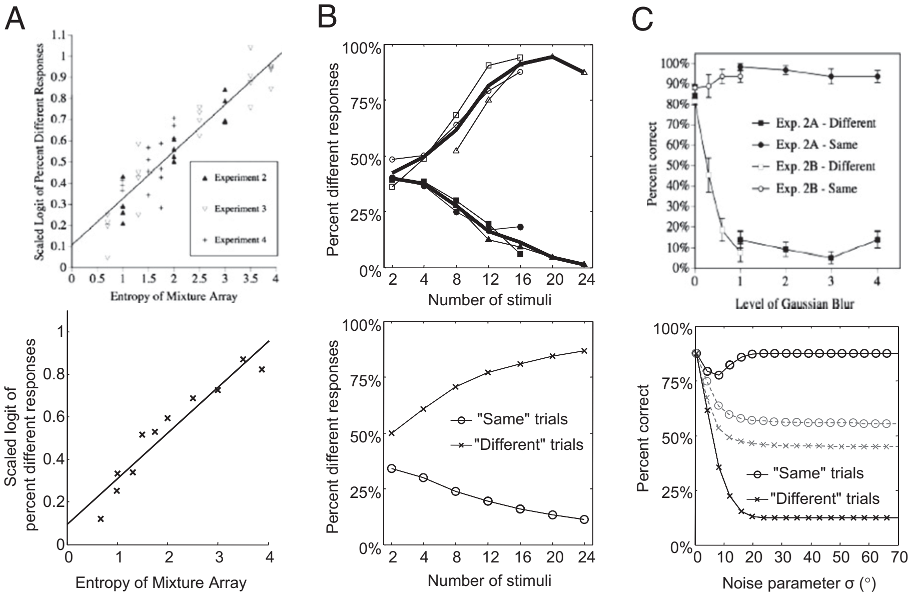

Fig. 5. Comparison between sameness judgment in animals and the optimal observer. (*A*) *Upper:* Data from Young and Wasserman (24). The proportion of "different" responses in pigeons correlates strongly with the entropy of the stimulus set. *Lower:* Results from an optimal-observer simulation. (*B*) *Upper:* Data replotted from Wasserman, Young, and Fagot (16). Thin lines represent individual experiments, the thick line their mean. Performance of baboons increases with set size. *Lower*: Results from an optimal-observer simulation. (*C*) *Upper:* Data from Wasserman and Young (28). Increasing the amount of blur in the stimulus set results in more "same" responses on "different" trials, but leaves the responses on "same" trials largely unaffected. *Lower:* Black, results from an optimal-observer simulation with a prior $p_{\mathrm{same}} = 0.6$ and a guessing rate of 0.25. Gray, same with $p_{\mathrm{same}} = 0.5$.

"same" and "different" trials, which improves discrimination accuracy. In other words, each additional stimulus provides additional evidence for "same" or "different." This stands in contrast to visual search (19) and change detection (18) with a single target, in which performance decreases with set size in part because every additional stimulus is a distractor and decreases the signal-to-noise-ratio.

Finally, we examined the effect of stimulus visibility on sameness judgment. Young and colleagues (20, 28) reported that decreasing stimulus visibility increased pigeons' tendency to respond "same" on "different" trials but had very little effect on responses on "same" trials (Fig. 5*C*, *Upper*). The optimal-observer model quantitatively accounts for the observed trends (Fig. 5*C*, *Lower*) by using a prior that slightly favors "same." Such a prior has recently been found in an unrelated same–different task in pigeons (29). As stimulus visibility decreases, this prior influences the animal's decision more strongly, pushing the proportion of "different" responses toward zero. Gray lines in Fig. 5*C* show the model prediction when the prior is equal to 0.5. We predict that when the prior for "same" is ≤0.5, performance on "same" trials will decrease with decreasing visibility. This is to be compared to Young et al. (20), who predict "same" performance will remain unaffected.

Altogether, these results show that the optimal-observer model, without modifications, accounts for several key findings from the animal cognition literature. Earlier models (20, 24) mapped a measure of variability within a display to a probability of responding "different." However, both the variability measure and the mapping were postulated rather than derived. The optimal-observer model, by contrast, derives the relationship between stimuli and response probabilities from the principle that the observer maximizes performance. It should be noted, however, that the picture stimuli used in the animal studies were complex and high-dimensional, whereas our model assumes one-dimensional stimuli. The similarities between data and model suggest that the strategy of comparing the variance of the measurements to a criterion, Eq. **1**, captures the essence of sameness judgment even for more complex stimuli.

Discussion

In a visual version of sameness judgment, we found that the optimal strategy consists of comparing the reliability-weighted variance of the measurements with a decision criterion that depends both on set size and on the reliabilities of the measurements. Human decisions are consistent with this strategy. In particular, observers use information about the reliabilities of individual items on each trial. These results provide a normative and mathematically precise footing for the study of sameness judgment and indicate that the notion of perception as near-optimal inference extends to perceptual relations.

The use of complex stimuli, such as photographs and clip art, might have hampered previous modeling of sameness judgment. Such stimuli are difficult to parameterize and therefore difficult to use as a basis for quantitative models. Here, we modeled sameness judgment in a single feature dimension. An important challenge that lies ahead is to extend the optimal-observer framework to more complex stimuli.

Sensory noise is not the only possible cause of uncertainty in sameness judgment. For example, judging whether a pair of dots or line segments belong to the same contour or to different contours is nontrivial even in the absence of sensory noise, because a given pair can be consistent with both hypotheses (30, 31). This is a form of ambiguity (see also ref. 32). Optimal-observer models have been successful in accounting for human behavior in such classification tasks (30–32). A logical extension of the present work would be to examine sameness judgment in the simultaneous presence of sensory noise and ambiguity. A particularly interesting manipulation in contour detection tasks would be to vary the reliability of the elements unpredictably from trial to trial, similar to experiment 2, to investigate whether knowledge of sensory uncertainty is optimally combined with knowledge of stimulus classes.

Human ability to efficiently take into account unpredictable local variations in reliability during sameness judgment supports the notion that the brain encodes entire likelihood functions over stimuli (33, 34) instead of only point estimates. This idea is already well known in the domain of cue combination (26). Cue combination, however, has a rather simple statistical structure. Many perceptual tasks, including sameness judgment, are more complex because of the presence of multiple relevant stimuli (spatial complexity) or because stimulus information must be integrated into an abstract categorical judgment (structural complexity). Only recently has weighting by reliability in tasks with such complexities begun to be explored (27). Weighting by reliability could potentially also be studied in high-level, cognitive forms of sameness judgment (22, 35).

Our findings constrain potential neural implementations of sameness judgment. Any candidate neural network would have to propagate information about the sensory uncertainties associated with individual items on a given trial and use this information near-optimally in subsequent computation. The framework of Poisson-like probabilistic population codes (34) has previously been used to solve this problem for cue combination (34), decision-making (36), and visual search (27). Applying it to sameness judgment could produce a network that approximates the correct posterior probability of sameness on each trial.

Materials and Methods

Experiment 1: Orientation. Subjects viewed the display on a 19″ LCD monitor from a distance of 60 cm. Background luminance was 30 cd/m^2. Stimuli consisted of a set of N gray (40 cd/m^2) oriented ellipses with an area of 0.5 deg^2 (we use "deg" to denote degrees of visual angle, and "°" for stimulus orientation or angle in the display). Ellipse eccentricity (elongation), which is defined as $(1 - b^2/a^2)^{0.5}$, where a and b are the lengths of the semimajor and semiminor axes, respectively, was fixed at 0.94. Set size was 2, 4, or 8. On each trial, the mean of the presented set of stimuli was randomly drawn from the range [0°, 180°], and the probability of a "same" trial was 0.5. Subjects were informed in advance about the way the stimuli were generated. The items were presented on an imaginary circle centered at fixation with a radius of 8 deg of visual angle. The position of the first stimulus was chosen at random on each trial; the other stimuli were placed in such a way that the angular distance between two adjacent stimuli was always 45°. This ensured that the average distance between two neighboring stimuli was constant across set sizes. Both the x and y position of each item was further subjected to a Gaussian random jitter with zero-mean and an SD of 0.25 deg. Each trial began with a presentation of the central fixation cross (500 ms), followed by the stimulus display (67 ms), followed by a blank screen until the subject responded. Feedback was given by coloring the fixation cross. The experiment consisted of three parts, each with a different set size (2, 4, or 8), presented in random order. Each part consisted of six blocks of 150 trials. The experiment was split into two sessions; the sessions took about 1 h each and were performed on different days. One author and seven paid, naïve subjects participated in the experiment. All stimuli were controlled using MATLAB (MathWorks) with Psychtoolbox. The color version of this experiment is described in *SI Appendix*.

Experiment 2. The stimuli, conditions, and procedure of experiment 2 were identical to those of experiment 1, except for the following differences. The set of stimuli consisted of six ellipses of fixed area. The angular distance between each two consecutive stimuli was 60 deg. The set of all stimuli was rotated around fixation over a random angle on each trial. Their reliability was controlled by ellipse eccentricity. Two values of eccentricity were used. The high value was always 0.94. To determine the low value, subjects performed five threshold measurement blocks before the main experiment; in these, all six stimuli had the same eccentricity, randomly chosen on each trial from 10 values linearly spaced between 0.5 and 0.94. Each of these blocks consisted of 150 trials. On the basis of these data, a psychometric curve was mapped out that related accuracy to ellipse eccentricity. A cumulative Gaussian function was fit to these data (accuracy as function of eccentricity), and the eccentricity at which a subject performed 70% correct was defined as their low eccentricity value. The first two threshold measurement blocks were considered practice and were not included in this analysis. In the testing blocks, only two values of eccentricity were used, as described in the main text. Ten observers each performed three LOW blocks, three HIGH blocks, and 12 MIXED blocks, presented in random order and each consisting of 150 trials. Two authors and eight paid, naïve subjects participated in this study.

ACKNOWLEDGMENTS. We thank David Krueger for his work on a pilot version of experiment 1 and Tony Wright for useful discussions. W.J.M. is supported by National Eye Institute Award R01EY020958. R.v.d.B. is supported by The Netherlands Organisation for Scientific Research. K.J. was supported by National Science Foundation Grants DMS-0817649 and DMS-1122094 and a Texas Advanced Research/Technology Program award.

1. James W (1890) *The Principles of Psychology* (Henry Holt, New York).
2. Gibson JJ (1950) *The Perception of the Visual World* (Houghton Mifflin, Boston).
3. Daehler MW, Bukatko D (1985) *Cognitive Development* (Knopf, New York).
4. Nissen HW, Blum JS, Blum RA (1948) Analysis of matching behavior in chimpanzee. *J Comp Physiol Psychol* 41:62–74.
5. French RS (1953) The discrimination of dot patterns as a function of number and average separation of dots. *J Exp Psychol* 46:1–9.
6. Rensink RA (2002) Change detection. *Annu Rev Psychol* 53:245–277.
7. Robinson EW (1933) A preliminary experiment on abstraction in a monkey. *J Comp Psychol* 16:231–236.
8. Bravo MJ, Nakayama K (1992) The role of attention in different visual-search tasks. *Percept Psychophys* 51:465–472.
9. Körding KP, et al. (2007) Causal inference in multisensory perception. *PLoS ONE* 2: e943.
10. Giurfa M, Zhang S, Jenett A, Menzel R, Srinivasan MV (2001) The concepts of 'sameness' and 'difference' in an insect. *Nature* 410:930–933.
11. Katz JS, Wright AA (2006) Same/different abstract-concept learning by pigeons. *J Exp Psychol Anim Behav Process* 32:80–86.
12. Pepperberg IM (1987) Acquisition of the same/different concept by an African Grey parrot (*Psittacus erithacus*): Learning with respect to categories of color, shape, and material. *Learn Behav* 15:423–432.
13. Mercado E, Killebrew DA, Pack AA, Mácha IV, IV, Herman LM (2000) Generalization of 'same-different' classification abilities in bottlenosed dolphins. *Behav Processes* 50:79–94.
14. Katz JS, Wright AA, Bachevalier J (2002) Mechanisms of same/different abstract-concept learning by rhesus monkeys (*Macaca mulatta*). *J Exp Psychol Anim Behav Process* 28:358–368.
15. Robinson JS (1955) The sameness-difference discrimination problem in chimpanzee. *J Comp Physiol Psychol* 48:195–197.
16. Wasserman EA, Young ME, Fagot J (2001) Effects of number of items on the baboon's discrimination of same from different visual displays. *Anim Cogn* 4:163–170.
17. Young ME, Wasserman EA, Garner KL (1997) Effects of number of items on the pigeon's discrimination of same from different visual displays. *J Exp Psychol Anim Behav Process* 23:491–501.
18. Phillips WA (1974) On the distinction between sensory storage and short-term visual memory. *Percept Psychophys* 16:283–290.
19. Estes WD, Taylor HA (1964) A detection method and probabilistic models for assessing information processing from brief visual displays. *Proc Natl Acad Sci USA* 52:446–454.
20. Young ME, Wasserman EA, Ellefson MR (2007) A theory of variability discrimination: Finding differences. *Psychon Bull Rev* 14:805–822.
21. Wilken P, Ma WJ (2004) A detection theory account of change detection. *J Vis* 4: 1120–1135.
22. Holyoak KJ, Thagard P (1995) *Mental Leaps: Analogy in Creative Thought* (MIT Press, Cambridge, MA).
23. Wasserman EA, Young ME, Cook RG (2004) Variability discrimination in humans and animals: Implications for adaptive action. *Am Psychol* 59:879–890.
24. Young ME, Wasserman EA (1997) Entropy detection by pigeons: Response to mixed visual displays after same-different discrimination training. *J Exp Psychol Anim Behav Process* 23:157–170.
25. Yuille AL, Bülthoff HH (1996) Bayesian decision theory and psychophysics. *Perception as Bayesian Inference*, eds Knill DC, Richards W (New York Univ Press, New York), pp 123–161.
26. Knill DC, Pouget A (2004) The Bayesian brain: The role of uncertainty in neural coding and computation. *Trends Neurosci* 27:712–719.
27. Ma WJ, Navalpakkam V, Beck JM, Berg R, Pouget A (2011) Behavior and neural basis of near-optimal visual search. *Nat Neurosci* 14:783–790.
28. Wasserman EA, Young ME (2010) Same-different discrimination: The keel and backbone of thought and reasoning. *J Exp Psychol Anim Behav Process* 36:3–22.
29. Wright AA, Katz JS, Ma WJ (2012) How to be proactive about interference: Lessons from animal memory. *Psych Sci*, in press.
30. Geisler WS, Perry JS (2009) Contour statistics in natural images: Grouping across occlusions. *Vis Neurosci* 26:109–121.
31. Feldman J (2001) Bayesian contour integration. *Percept Psychophys* 63:1171–1182.
32. Feldman J, Tremoulet PD (2006) Individuation of visual objects over time. *Cognition* 99:131–165.
33. Pouget A, Dayan P, Zemel RS (2003) Inference and computation with population codes. *Annu Rev Neurosci* 26:381–410.
34. Ma WJ, Beck JM, Latham PE, Pouget A (2006) Bayesian inference with probabilistic population codes. *Nat Neurosci* 9:1432–1438.
35. Kroger JK, Holyoak KJ, Hummel JE (2004) Varieties of sameness: The impact of relational complexity on perceptual comparisons. *Cogn Sci* 28:335–358.
36. Beck JM, et al. (2008) Bayesian decision-making with probabilistic population codes. *Neuron* 60:1142–1145.