Published in final edited form as:
J Vis. ; 9(1): 6.1–619. doi:10.1167/9.1.6.

The Mixture of Bernoulli Experts: A theory to quantify reliance on cues in dichotomous perceptual decisions

Benjamin T. Backus
Department of Vision Sciences, State University of New York, State College of Optometry, New York, NY, USA

Abstract

The appearances of perceptually bistable stimuli can by definition be reported with confidence, so these stimuli may be useful to investigate how visual cues are learned and combined to construct visual appearance. However, interpreting experimental data (percent of trials seen one way or the other) requires a theoretically motivated measure of cue effectiveness. Here we describe a simple Bayesian theory for dichotomous perceptual decisions: the Mixture of Bernoulli Experts or MBE. In this theory, a cue's subjective reliability is the product of a weight and an estimate of the cue's ecological validity. The theory (1) justifies the use of probit analysis to measure the system's reliance on a cue and (2) enables hypothesis testing. To illustrate, we used apparent 3D rotation direction in perceptually ambiguous Necker cube movies to test whether the visual system relied on a newly recruited cue (position of the stimulus within the visual field) to the same extent when a long-trusted cue (binocular disparity) was present or not present in the display. For six trainees, reliance on the newly recruited cue was similar whether or not the long-trusted cue was present, suggesting that the visual system assumed the new cue to be conditionally independent.

Keywords

cue combination; cue recruitment; cue learning; bistability; ambiguous figure; perceptual dichotomy; appearance; sensory fusion; machine learning; Bayes; naive Bayes; Bayes rule

Introduction

The process by which retinal images become the visual percepts we "see" is still a mysterious one. It is nevertheless clear that theory can be usefully applied to predict perceptual experience (or appearance or qualia) when a person opens her eyes and looks at a particular stimulus. Recent work describes appearance as resulting from the construction of representations of the environment by processes that are near-optimal in their ability to use measured visual signals (e.g., Backus & Banks, 1999; Brainard & Freeman, 1997; Feldman, 2001; Feldman & Tremoulet, 2006; Geisler & Kersten, 2002; Geisler, Perry, Super, & Gallogly, 2001; Hillis, Watt, Landy, & Banks, 2004; Hogervorst & Eagle, 1998; Kersten, Mamassian, & Yuille, 2004; Knill & Richards, 1996; Weiss, Simoncelli, & Adelson, 2002). This approach has its origin in older ideas about the probabilistic nature of visual information (Brunswik, 1956; Fechner, 1860; Helmholtz, 1910/1925; Hochberg & Krantz, 2004). In these theories, estimated

Corresponding author: Benjamin Backus. Email: bbackus@sunyopt.edu. Address: Vision Sciences, SUNY Optometry, 33 W. 42nd St., New York, NY 10036, USA.

Commercial relationships: none.

scene properties are manifested consciously within visual percepts as perceptual *attributes* such as perceived velocities, surface slants, surface colors, contour groupings, object identities, and so on. The fusion of sensory data to estimate scene parameters has received particular attention, because theory makes testable predictions about how various cues (each of which is a statistic computed on the stimulus) are combined with each other and with prior belief to construct a perceptual attribute that represents a scene property.

Cue recruitment experiments have recently demonstrated that the visual system is capable of learning to utilize new cues during the construction of appearance (Haijiang, Saunders, Stone, & Backus, 2006). It is not clear, however, how to measure the strength of these effects. A theory to quantify the learning would make it possible to use bistable stimuli in quantitative tests of hypotheses about cue recruitment and cue combination. The purpose of this paper is to describe such a theory, called the Mixture of Bernoulli Experts (MBE). This theory justifies the use of probit analysis to quantify cue effectiveness during dichotomous perceptual decisions.

Bistable stimuli and the study of appearance

Appearance is probably an important mediator of behavior. After the visual system decides how to represent the local environment as appearance, a second round of decisions that depend on appearances becomes possible: what deliberate actions to take (Brunswik, 1956; Hebb, 1949), what to remember about the world (Barlow, 1990), or how one should describe things to another person. It is therefore imperative to study appearance *per se* and imperative to study how learning causes changes in appearance. However, tools for this study are limited. A "perceptual learning" experiment that measures an improvement in discrimination performance can tell us very little about how the stimulus looked to the observer. Indeed, observers often report being unsure how a stimulus looks at threshold. In perceptual matching and perceptual nulling experiments, observers are instructed (and are evidently able) to minimize differences in appearance between two stimuli (or between a stimulus and an internal standard in the case of nulling) but appearance itself is not revealed by these experiments.

Instead, verbal report must be elicited if the experimenter wants to know what the stimulus looks like to the observer. But classic psychophysical procedures were not designed to maximize the reliability of verbal reports about appearance. Reliable verbal report about appearance *requires* that the observer have high confidence about the appearance of the stimulus as constructed by his or her perceptual system. Stimuli that evoke a truly dichotomous perceptual response solve this problem and for that reason are a useful tool in the perceptual scientists' repertoire.

This advantage is particularly important to the study of cue recruitment. A suprathreshold change to a newly recruited cue may be needed to achieve a just-noticeable effect on a particular perceptual attribute. If the newly recruited cue causes some other perceptual attribute to vary —as is almost always the case—then the observer will find it difficult if not impossible to follow the experimenter's instruction to use only the trained attribute when responding. The other perceptual attribute, which was controlled by the cue prior to training in the experiment, will be difficult for the observer to ignore.

Procedures that use perceptually bistable stimuli are thus potentially very useful: if a perceptual attribute takes on exactly one of two well-defined values on each trial, it is trivial for the observer to follow the experimenter's instruction to report the form of the attribute. We are therefore in need of theory to quantify the effects of stimulus manipulations on perceptually bistable stimuli. Percent of trials seen one way or the other is the direct dependent measure, but we would like if possible to find a measure that is theoretically motivated and, if possible, mathematically better behaved.

Use of probit analysis to quantify strength of evidence in a perceptually bistable stimulus

Dosher, Sperling, and Wurst (1986) used probit analysis to quantify the relative effectiveness of three visual cues—binocular disparity, perspective, and proximity luminance—that biased the apparent direction of rotation of a Necker cube. When the "strength of evidence" from each cue was described as a *z*-score—also known as probit units or NEDs (normal equivalent deviations, Gaddum, 1933)—their effects were found to be additive. The work we describe here builds on that of Dosher et al. (1986) in two ways. First, we show that probit analysis is appropriate because the *z*-scores can be interpreted as belief terms that correspond to likelihoods for the probability parameter in a binomial decision model. Second, we use the model to test the proposition that the visual system's trust in a newly recruited cue is the same whether or not a long-trusted cue is present in the stimulus. In the Bayesian statistical framework implied by the model, additive cue effects result when the visual system believes that cues are conditionally independent.

These findings illustrate that perceptually dichotomous stimuli can be understood, and therefore used to study cue combination, within the established Bayesian framework that has proved useful to study continuously valued perceptual attributes. They also suggest that by default the visual system assumes a newly discovered visual cue to be conditionally independent of the long-trusted cues from which the new cue's meaning was inferred.

Previous approaches to the problem

Bayesian formulations that were developed to model continuously valued perceptual attributes might be adapted for use with binary attributes. For example, the resolution of perceptual bistability on a given trial might be treated as a degenerate case of model selection (Knill, 2003), where the two "models" are simply constant functions. However this does not really solve the problem because it does not specify how to choose between the two models on a given trial.

Two Bayesian models were previously developed to explain specific examples of perceptual bistability: line drawings of egg/saddle shapes (Mamassian & Landy, 1998) and rendered 3D ridges/valleys (Mamassian & Landy, 2001). These domain-specific models described the effects of priors and cues on appearance but they did not provide a general framework for the analysis of perceptually bistable stimuli. These models also differ from MBE in their use of a "noncommitting" decision rule: variability in appearance was modeled not as due to a noise process, but rather to probability matching. In other words, the posterior was understood to directly specify the probability of seeing the stimulus one way or the other. This may reflect a deep truth about perception, but our suspicion is that it confuses decision noise with log likelihood (see Appendix A).

Another Bayesian model captured the bimodality of (continuously valued) apparent slant in stimuli that contained large cue conflicts, using changes in cue weighting over time (van Ee, Adams, & Mamassian, 2003). Our approach develops a simpler theory that applies specifically to dichotomous perceptual decisions. This choice allows us to use simpler mathematics and lends itself naturally to testing hypotheses about cue combination and cue recruitment. It applies to any perceptually bistable stimulus presented in a brief, discrete trial. MBE does not predict the time course of spontaneous alternation in appearance (Attneave, 1971; Leopold & Logothetis, 1999; Necker, 1832).

Combinations of experts

Suppose that each of two experts, using different cues, is asked to estimate the probability that a binary property of the world is in State 1 (as opposed to State 0). Suppose they give estimates of, say, $p = 0.4$ and $q = 0.8$, respectively. How should we combine these estimates into a single

subjective probability? If we believe the experts are well calibrated and have independent information, we might use the multiplicative rule: combined probability = $pq / [pq + (1 - p)(1 - q)] \approx 0.73$. However, the experts' estimates may not be independent (their cues may co-vary for example) or one of the experts may understand better than the other how to interpret its cue. It could turn out that Expert 1 is much better than Expert 2, in which case State 0 is more probable than State 1. To use these experts optimally, a decision maker would need full knowledge of the joint probability between the binary property, the probability estimates made by Expert 1, and the probability estimates made by Expert 2 (Smith & von Winterfeldt, 2004).

Morris (1983) illustrates how the combining of probabilities depends on specific circumstance. Suppose two experts have experience of the same biased coin, and both give a subjective probability of 0.6 that the next toss will be heads. In this case our best combined probability is clearly 0.6, not the probability given by the multiplicative rule (which is 0.69). A less clear-cut case is estimating the probability of rain. If two weather forecasters both estimate the chance of rain at 60%, is our best estimate 60%, in agreement with both forecasters, or is it higher? As with our example above, the answer will depend on circumstance.

Thus, a variety of approaches have been proposed for combining experts' estimates of probability in the absence of full information (Clemen & Winkler, 1999; Hummel & Landy, 1988; Shafer, 1976, 1981). MBE is an adaptation of an approach that was first developed by Morris (1983) and Winkler (1968), which estimates a Bernoulli probability as a weighted average of experts' probability estimates (Clemen & Winkler, 1999). This idea will be new to many readers (as it was to the author) and MBE includes a term for decision noise that is new. We therefore include in the next section the author's derivation of MBE that was developed independently in the context of explaining dichotomous perceptual decisions.

Without fully modeling the joint probability, MBE allows that some cues are better than others, and that the "expert" responsible for using a given cue in one individual's perceptual system may be better than the "expert" who uses that cue in another individual's system. The model allows us to reduce these two factors to a single number that no longer distinguishes between a good cue and a good expert. In the context of the model, the goodness of a cue (its dependence on the binary property of the world) is the cue's *ecological validity*; the estimated quality of the expert is the cue's *weight*; and the single number that combines estimated ecological validity and weight is the cue's *moment* or *subjective reliability*.

By way of a final example, consider a person who goes camping for the first time in a new country and who notices red ants that bite. After 10 minutes, having seen four ants on his skin and being bitten once, the camper estimates that red ants bite $\hat{\theta} = 0.25$ of the time. The camper cannot be sure—perhaps the ants actually bite 1/10 or 1/2 of the time on average. After a week, he may have seen hundreds of ants on his skin and been bitten 1/4 of the time. Again his best estimate is $\hat{\theta} = 0.25$, but this estimate is now much more precise. To capture this difference one can model belief about θ as a probability density function of the possible values for θ from 0 to 1. Doing something similar for each of several cues enables us to combine the cues in a sensible fashion based on their weights. It also allows us to update an existing estimate of θ in light of a new cue to get the same estimate of θ that would be obtained if the cues were present in the first place ("external Bayesianity," e.g., Clemen & Winkler, 1999).

Linear combinations of estimators (such as the weighted average in MBE) have computationally robust optimality properties (Foster & Vohra, 1999) and have proven remarkably successful in real-world applications (Clemen & Winkler, 1999; Kohavi, 1996).

Theory for a binary perceptual decision from multiple cues

The theory describes the single perceptual outcome of a given trial in a real psychophysical experiment. We assume that the stimulus and perceptual system jointly specify exactly two possible outcomes for the state of the perceptual attribute $\boldsymbol{A}$, so that $\boldsymbol{A}$ will take on either A_0 or A_1 as its value. A_0 and A_1 stand for two perceptual representations, corresponding to two states that the world might actually be in. Thus, the perceiver believes that she will be either correct or incorrect depending on whether the correct value of $\boldsymbol{A}$ is selected (it is immaterial at this point whether she receives feedback about the correctness of her choice).

Three factors influence the decision between A_0 and A_1:

1. the system's prior belief, conditioned on all factors other than noise and the cues;
2. a noise process that contributes differently from trial to trial, which captures all factors besides the prior and the cues (including the state of any endogenous process that prefers one form or the other at different times, unmodeled sequential order effects due to recent trials, effects of counterbalanced stimulus properties that were not modeled explicitly, and any uncontrolled variables); and
3. the cues, each of which is assumed to depend independently on the state of the world. If two signals are not conditionally independent then to participate in the model, they must be combined to create a single, new cue that captures their interaction, in order to participate as an "expert" in MBE.

What is a "cue" in MBE?

For the sake of precision, our use of the word "cue" differs from conventional usage, which is ambiguous. In our hands a signal (i.e., a statistic computable from the optic array or other sensory input) is not a "cue" by virtue of its perceptual effect, but rather by virtue of being dependent upon, and thus informative about, a property of the world. Thus the system "fails to use a cue" if it fails to make use of a predictive signal from the environment; an informative signal is still a cue even if the perceptual system fails to exploit it.

Conversely, we discourage use of the word "cue" in one of its particular traditional meanings, as a signal that affects appearance or behavior. Occasionally the perceptual system uses a signal to construct appearance as if it had ecological validity when in fact it does not (e.g., Ho, Landy, & Maloney, 2008; Ho, Maloney, & Landy, 2007). If an irrelevant signal participates in the construction of a perceptual attribute we will say that signal is "treated by the system as though it were a cue" but not that it *is* a cue. In short, whether a signal is a cue depends on the statistical properties of the environment, not on the properties of cognitive mechanisms. Informally we will sometimes also use the word "cue" to speak of a signal within the perceptual system, constructed by it from the sense data, but properly we should call this a "cue measurement."

In the laboratory, an "artificial cue" can be created by putting some signal into correlation with cues that are informative in the natural environment (and trusted by the perceptual system for that reason). This new "cue" may (or may not) be recruited by the perceptual system for purposes of constructing appearance. Just as the world supplies a set of signals with statistical relationships from which it can be known, the experimenter may create artificial statistical relationships from which his artificial world can be known. Our theoretical framework specifies that the perceptual system assumes the existence of a real world and tries to represent it; we therefore assume that it decides whether or not to use artificial cues similarly to how it decides whether or not to use natural cues. The theory does not address how the system decides which cues to use in the first place (which is the problem of *feature selection* in machine learning; e.g., Dietterich, 2002).

Bayesian model for cue combination

We model the perceptual decision as a guess about the outcome of a Bernoulli trial. In order to make its guess, the system estimates θ, the Bernoulli probability, from the expected value of a posterior probability function $\pi(\theta)$ on the interval [0, 1] for θ. The system chooses A_0 or A_1 as the percept according to a decision rule: if the expected value of $\pi(\theta)$ is less than 0.5, the system chooses A_0, and if it is greater than or equal to 0.5 it chooses A_1. We use the expected value because it is mathematically tractable and because if $\pi(\theta)$ has the form of a beta distribution, then the expected values is on the same side of 0.5 as other measures of central tendency. The cutoff of 0.5 is chosen for symmetry; it also maximizes the probability of making a correct guess when estimates of θ are unbiased.

To estimate θ the system must utilize prior belief and the cues, each of which supplies its own separate "expert" estimate of θ. As we shall see, a sensible way to combine these estimates is to update a beta distribution to determine the posterior for θ. To this end, n_i quantifies the *weight* of the *i*th cue ($i = 1,2, \ldots, N$) as an "equivalent number of observations from a binomial experiment". The *i*th cue therefore supplies a pair of numbers (θ_i, n_i) to be used as evidence in the perceptual decision. θ_i is a probability, so it is a real number between 0 and 1, inclusive, and n_i is a nonnegative (possibly non-integer) real number. Thus the weight of a cue may be greater than 1 and weights need not sum to 1.

As a convention we will arbitrarily adopt $i = 0$ for the prior and $i = -1$ for decision noise (discussed later). Cues 1 through N are based on sensory evidence. There are thus $N + 2$ factors or experts that contribute to the decision at a given time, corresponding to $i = -1, 0, 1, 2, \ldots, N$.

The beta distribution is the conjugate prior for the binomial distribution, meaning that if one estimates a fixed Bernoulli probability θ by conducting a binomial experiment, and one's prior density function for θ is a beta distribution, then the posterior density is also a beta distribution (Gelman, Carlin, Stern, & Rubin, 1995). If the prior is beta (α_0, β_0) and the new binomial experiment yields x successes and y failures on a total of $n = x + y$ trials, then the posterior is beta (α_1, β_1) where $\alpha_1 = \alpha_0 + x$ and $\beta_1 = \beta_0 + y$. The expected value of beta(α, β) estimates θ and is equal to $\alpha/(\alpha + \beta)$. Thus, after the experiment, the estimate of θ will have moved from $\alpha_0/(\alpha_0 + \beta_0)$ to $(\alpha_0 + x) / (\alpha_0 + \beta_0 + n)$. This fact makes it natural to think of the prior as being equivalent to a previous binomial experiment that had α_0 successes and β_0 failures in $n_0 = \alpha_0 + \beta_0$ total trials; adding more trials by doing a new binomial experiment simply increases the total number of successes and/or failures in the two binomial experiments combined.

Our choice of the beta distribution for $\pi(\theta)$ in the model means that $E(\pi(\theta))$ is conceptually equivalent to an estimate of θ for a binomial experiment. The prior is beta(α_0, β_0) where $\alpha_0 = n_0\theta_0$ and $\beta_0 = n_0(1 - \theta_0)$. When using the model to describe variation in responding during a psychophysical experiment, the prior provides a constant bias that favors A_0, A_1, or neither depending on whether θ_0 is less than, greater than, or equal to 0.5, respectively. The noise term modifies the estimate of θ by adding $n_{-1}\theta_{-1}$ to α_0 and $n_{-1}(1 - \theta_{-1})$ to β_0. On each trial of the psychophysical experiment $\theta_{-1} \in [0,1]$ is a random variable that depends on model-specific assumptions, with weight n_{-1}. In the current version of the model we assume $n_{-1} = 10$ for convenience to achieve a standard normal noise term (see Noise term section below). Finally, each of the cues (based on sense data) contributes its own additional biasing effect in similar fashion.

The noise, prior, and cues thus combine to produce a posterior of the form:

$$\pi(\theta)=\text{beta}\left(\sum_{i=-1}^{N} n_i\theta_i, \sum_{i=-1}^{N} n_i(1-\theta_i)\right), \tag{1}$$

and the final estimate of θ (namely $\hat{\theta}$) is given by what we call the *mixture of Bernoulli experts* (*MBE*) equation:

$$\widehat{\theta}=E(\pi(\theta))=\sum n_i\theta_i/\sum n_i. \tag{2}$$

Equation 2 shows $\hat{\theta}$ to be a *probability mixture model* and gives its *probability mass function*. The use of MBE to recover cue weights is tantamount to measuring a mixture of beta distributions.

Figure 1 illustrates the model for a toy example using no noise ($n_{-1} = 0$), Jeffrey's prior ($\theta_0 = 0.5$, $n_0 = 1$), and two cues [(θ_1, n_1 = (1/3, 45) and (θ_2, n_2) = (0.8, 10)]. The prior is unbiased, the first cue favors A_0 (because $\theta_1 < 0.5$), and the second cue favors A_1. According to Equation 2 the value of $\hat{\theta}$ will be 23.5/56 < 0.5 so the model predicts that A_0 will be perceived.

Model properties and operational choices—Several observations and clarification of operational choices are in order:

1. From Equations 1 and 2 it can be seen that the effectiveness of a cue is not determined by serial processing order. The $N + 2$ factors are treated as having simultaneous effects. In perception, however, the system's choice between the two forms of a perceptual attribute may depend critically on relative cue onset times and this effect can be very different from one individual to another (Dosher et al., 1986). Thus, MBE can be used to quantify the system's reliance on different cues within a given stimulus for purposes of comparing stimuli but does not describe dynamic aspects of the cue combination process.
2. In some psychophysical experiments the system's error in measuring some *i*th cue will be a significant source of noise. One can model this noise as affecting the system's estimate of θ_i or n_i or both, thereby turning them into random variables. That approach may be useful to generate simulated data for experiments in which cue reliability is manipulated. For simplicity, however, and in contrast with models in which a cue's weight is proportional to its trial-to-trial reliability (Backus & Banks, 1999), it will sometimes be a defensible modeling choice to collect this source of noise into the noise term (as we have done to model data from the Test below).
3. The decision rule in MBE is deterministic: is $\hat{\theta} \geq 0.5$? In the Discussion section we consider probability matching as an alternative decision rule, in which case the decision process itself is probabilistic.
4. Mathematically, it is impossible in the model to increase the expected frequency of outcome A_1 by adding an additional cue for which θ_{N+1} is 0.5 or less. One can cause $\hat{\theta}$ to move closer to 0.5 on average (across trials of a psychophysical experiment), but $\hat{\theta}$ will not exceed 0.5 on a greater fraction of trials, assuming fixed noise. This result follows from the construction of $\hat{\theta}$ as a weighted average.
5. The model allows for two cues each to completely control appearance when the other is absent, yet for one of them to dominate when both are present in conflict with each other (so long as $\hat{\theta}$ is to the same side of 0.5 on all trials).

6. θ_i is the probability of A_1 given the *i*th cue. We do not specify whether the *i*th cue represents information available in the optic array, on the retina, or in a later neural representation in the visual system (in which case θ_i must be closer to 0.5, given measurement noise). This distinction corresponds to a choice of where to put the "ideal observer" (Geisler, 1989) when doing a particular analysis; the modeler may choose whichever meaning of "cue" seems most appropriate to circumstances at hand.

7. θ_i depends on the statistics of the environment. What one knows about the environment is therefore paramount: the true value of θ_i can become increasingly extreme (closer to 0 or 1) as one differentiates environmental contexts. Consider the (crossed) binocular disparity between two objects in a scene, used as a cue to decide which object is closer. When viewing nearby objects directly, this cue has high ecological validity and θ_i is close to 1. But when viewing distant objects through distorted window glass (which corrupts disparity) θ_i is only slightly above 0.5. If the system does not distinguish these two environments from one another, then θ_i will have an intermediate value that depends on the proportion of time the system actually spends in each environment.

8. For convenience we use the symbol θ_i both to indicate a real probability and the system's belief about the value of that probability. A *well-calibrated observer* is one who estimates θ_i correctly for a given environmental restriction.

9. By identifying cue weights with numbers of observations in a binomial experiment, we have been able to model cue strengths, and cue combination, within a framework of Bayesian inference. These hypothetical "observations" should not be confused with the observations of real stimuli made by the visual system during a cue recruitment experiment in the laboratory. Clearly, the latter sort of observation can cause changes over time in a cue's weight. However, MBE by itself does not provide a satisfactory model for cue recruitment because—at a minimum—an additional parameter is necessary to capture the fact that some cues are learned more quickly than others (see MBE and Bayes rule section).

Noise term—Trial-to-trial variation in the appearance of a perceptually bistable stimulus cannot be attributed completely to noise in the measurement of trusted cues. Other influences include intended variations in the stimulus from trial to trial that are not modeled explicitly, appearance on recent trials (Long, Toppino, & Kostenbauder, 1983), and non-deterministic endogenous processes that actively resolve the competition between the two forms on a given trial. To capture these sources of perceptual variation we use a separate term for decision noise, which we represent as a −1st contributing factor, (θ_{-1}, n_{-1}).

Because there may be many factors that contribute to the noise, we suggest modeling the "standard noise" term $n_{-1}\theta_{-1}$ as being normally distributed. It will be convenient in the next section if the quantity $(n_{-1})(2\theta_{-1} - 1)$ is distributed as the standard unit normal. We achieve this by arbitrarily assuming that θ_{-1} on each trial of the psychophysical experiment is equal to a random variable $\boldsymbol{\theta}_{-1}$ that has normal distribution, with mean 0.5 and standard deviation 0.05, and that $n_{-1} = 10$.

The noise contributes variance to the decision variable [right side of Equation 2]. If part of this variance becomes understood by the experimenter, the relative weight of the noise term in the model will fall, as part of the variation in appearance becomes explained by variation in a new $(N + 1)$st factor or cue. For example, the experimenter might discover that recent trial history, or some variable aspect of the stimulus, contributes to the perceptual decision from trial to trial[1]. As in signal detection theory, the natural unit to use to quantify effects is a standard

deviation of the noise, also known as a *z*-score unit or *NED* (normal equivalent deviation, Gaddum, 1933).

Simplification by use of moments

In the model, the *i*th cue can be effective either if n_i is large relative to the noise (with θ_i possibly deviating only a small amount from 0.5) or because θ_i is close to 0 or 1 (with n_i not necessarily large relative to noise). Whether the perceptual system makes a distinction between n_i and θ_i is not known. In any case, measuring the effectiveness of a cue—its ability to control the percept—does not require specifying the subjective reliability (n_i) and subjective probability (θ_i) separately. Accordingly, a simplified version of the model will often be preferred so we develop that simplified model here.

Rather than asking whether $\hat{\theta} \geq 0.5$, we ask equivalently whether $2\hat{\theta} - 1 \geq 0$. Using Equation 2 we construct a new decision variable **M**, with $\mathbf{M} \geq 0$ being the criterion for choosing A_1 over A_0:

$$\begin{aligned} \mathbf{M} &= (2\widehat{\theta} - 1)\sum n_i \\ &= \sum n_i(2\widehat{\theta_i} - 1), \end{aligned} \tag{3}$$

$$\begin{aligned} &= \sum n_i\widehat{\rho}_i \\ &= \sum m_i. \end{aligned} \tag{4}$$

The quantity $\rho_i = 2(\theta_i - 1)$ we define to be the *i*th cue's *ecological validity*. It should be noted that in our usage, this term is a rescaled probability that applies to a single value of a cue or to a difference of effect between two values of a cue, as predictors of state for a binary property of the world. This differs from Brunswik's (1956) definition of the term, which was the correlation between the cue's values and a (typically) continuous property of the world. We justify our use on the grounds that Brunswik's use and our new use both quantify the predictive value of a cue and because probability provides a better description of the relationship between cue and property of the world in the case of a binary property. Like correlation, ρ_i has range $[-1, 1]$. To say that a cue has an ecological validity of 0 or 1 is to say the same thing whether one is referring to a correlation or to ρ_i as defined in MBE.

In classical mechanics, a *moment* is the product of a tangential force and lever-arm length. This is a valuable definition because it captures the fact that different combinations of force and arm length are equally effective, and because several moments in a single system are additive. For both reasons, the addend $m_i = n_i\rho_i$ can be seen as analogous to a moment in classical mechanics, with ρ_i playing the role of force and n_i playing the role of distance from the axis of rotation. A "moment" in classical mechanics, as here, is not a normalized quantity (unlike a "moment" in statistics). Moments in classical mechanics are additive as are $\{m_i\}$ in Equation 4. Accordingly we suggest the term *moment* to describe m_i, which is the weight of evidence, be it negative or positive, in favor of A_1. The *grand moment* is **M**, with a range of $[-\infty, +\infty]$.

[1]When variance comes to be understood and is moved out of the noise term, the modeler has two choices. The model can be renormalized so that the remaining unexplained noise is once again of the form $(\boldsymbol{\theta}_{-1}, n_{-1}) = (\boldsymbol{N}(0.5, 0.05^2), 10)$, in which case the subjective reliabilities $n_0, n_1, n_2, \ldots, n_m$ would all increase by the same coefficient. Equivalently, the original noise could be used as a reference point, with the new n_{-1} being less than 10. The point is raised here purely as a point of understanding to illustrate the relationship between the subjective reliabilities of the noise and cues.

Equation 4 describes the perceptual decision variable, M, as a simple sum of moments corresponding, as before, to the noise ($i = -1$), the prior ($i = 0$), and measured cues ($i = 1, 2, \ldots, N$). The moment for the "standard noise" term m_{-1} is distributed as a standard normal random variable by construction. Equation 4 now provides a theoretical justification for the use of probit analysis to quantify the effectiveness of the prior and cues, as quantified by their respective moments.[2]

A cue's *subjective reliability* can be defined as identical to its moment (see Contributors to subjective reliability section).

Relationship to theory of signal detection

The theory of signal detection/discrimination (TSD, Coombs, Dawes, & Tversky, 1970; Green & Swets, 1966) is mathematically closely related to the theory presented here. Indeed, one can measure the difference in subjective reliability for two levels of a cue as a unit-less d' value by (a) assuming Gaussian decision noise, (b) presenting stimuli that contain the two levels of the cue while holding other stimulus properties constant, and (c) asking the observer on each trial whether she sees A_0 or A_1. When the stimulus is seen in agreement with the cue (as decided by the experimenter) a hit or correct rejection occurs; when seen opposite, a false alarm or miss occurs. When the noise term in Equation 4 is assumed to be standard normal, probit analysis will estimate the difference in a cue's moment for two levels of the cue, in units of standard deviation of the noise. This difference is d' for that cue. The subjective reliability of prior, m_0, describes the perceptual system's bias and its additive inverse, $-m_0$, is comparable to the criterion β in TSD.

Change in d' per trial (or training session) has previously been used as a measure of the rate at which performance of the entire observer improves during learning to discriminate or identify (Fine & Jacobs, 2002). A mathematically identical measure can be used to describe cue recruitment data, to chart change over time in the utilization of a cue in a dichotomous perceptual decision. Of course, that same cue usually already participates in the construction of some other perceptual attribute. In that case stimuli containing two values of the cue will be easily discriminated despite threshold level effects on the new attribute under study. Thus, Haijiang et al. (2006) found that upward vs. downward translation was recruited by trainees' visual systems as a cue that specified the 3D rotation direction of a Necker cube. After training, this cue was still very effective for controlling apparent translation direction, so stimuli were easily discriminable, but it was only somewhat effective for controlling apparent 3D rotation, which was the attribute under study.

It therefore seems prudent to adopt a name other than d' to use when measuring a cue's effect —or several cues' combined effect—on the appearance of a specific perceptual attribute. We suggest d^* to be measured (like d') in NEDs. d^* is measured by estimating M for two levels of a cue and taking the difference. To avoid confusion we recommend that "d^*" not be used except after making a convincing argument that the observer's responses during the experiment were mediated by the perceptual attribute under study.

For reference, Table 1 reviews the symbols and their meanings for the theory up through the definition of d^*.

[2]Mathematically **M** is the expected number of successes minus the expected number of failures, for a binomial experiment consisting of $\sum n_i$ trials with Bernoulli probability $\theta = \hat{\theta}$. For quantifying moments, this mathematical fact is not particularly useful, because the moments are measured in NEDs (relative to the noise). See Discussion section however for use of this fact in consideration of Bayes Rule as a learning rule during cue recruitment.

Test of conditional independence for a newly recruited visual cue

In MBE, a cue's effect on appearance is modeled as a consequence of the system's belief that a cue has ecological validity. Binary response data from experiments with perceptually bistable stimuli can be used to quantify the strength of belief in order to test specific hypotheses about how cues are learned or combined. We illustrate this by testing one such hypothesis: that the system, when it recruits a new cue, treats the new cue as though its value depends directly on (and is therefore directly informative about) some property of the world—as opposed to depending on (and being informative about) the value of the long-trusted cues with which it covaried during training. In other words, we hypothesize that the system acts by default as a naive Bayes classifier (Backus & Haijiang, 2007; Lewis, 1998) with the new cue being assumed conditionally independent of the long-trusted cues from which the new cues' meaning was inferred.

To test this hypothesis we assert that conditional independence predicts that the new cue will be equally effective whether or not the long-trusted cues are also present within the test display. In MBE, each belief term is quantified relative to the noise, not relative to the belief terms for other cues. If a newly recruited cue has a meaning to the perceptual system that is independent of the meanings of other cues, then its subjective reliability as measured in NEDs should not depend on the values of the other cues, or indeed, whether they are present in the display. Thus, the presence of long-trusted cues should neither facilitate nor diminish the new cue's effect on appearance. Although it is conceivable that the new cue could show equal effectiveness despite strong interaction with long-trusted cues, such an interaction would be highly coincidental and thus unlikely. A more likely interaction would be that the new cue is believed by the system to be informative about the value of the long-trusted cues. In that case, we might reasonably predict that removing the long-trusted cues from the display would increase the effect of the new cue, because the value of the long-trusted cue could then be known only from the new cue, not also from the stimulus.

Thus, Figure 2 uses a causal Bayes net to diagram two possible changes in belief that might occur during a cue recruitment experiment. Before the experiment (left graph) the system implicitly believes that some internal state variable **T** depends on a binary property **A** of the world. This make **T** informative about the state of **A** and so **T** is trusted as a cue for inferring the state of **A**. When a second internal variable **N** is put experimentally into covariation with **T** (symbolized for two-point training by the scatterplot icon), **N** is recruited as a cue for inferring the state of **A**.

The middle and right graphs of Figure 2 show two possible inferences about the cause of variation in **N** that could explain the system's new use of **N** to estimate **A**. In the middle graph, **N** depends on **T**, so **N** and **T** are not conditionally independent. **N** is directly informative about **T**, but only indirectly informative about **A**. We might therefore expect **N** to have a larger impact on the system's inferences about **A** when the stimulus does not support separate direct measurement of **T** (indeed, the middle graph shows **T** to be a d-separator of **N** and **T**). Some evidence against this net is already at hand because a new cue can be effective on its own (Haijiang et al., 2006).

In the right graph, **N** depends on **A**. Now $\Pr(\mathbf{N}|\mathbf{A},\mathbf{T}) = \Pr(\mathbf{N}|\mathbf{A})$, i.e., **T** and **N** are independent when conditioned on **A**. As a result **A** can be estimated from **N** at least as well as **T** can be estimated from **N** and we expect experimental manipulation of **N** to have equal impact on inferences about **A** whether or not **T** can be measured in the stimulus.

Critically, MBE justifies our use of probit analysis to test whether a new cue has "equal effectiveness" in stimuli that do or do not contain the long-trusted cue, respectively. The test is simply whether the new cue's moment (effect size in NEDs) is different in stimuli that do

or do not contain the long-trusted cue. If the new cue has the same effect for both stimuli, it supports the hypothesis.

The rotating wire frame (Necker) cube is a model system for studying the recruitment of new visual cues (Haijiang et al., 2006). Movies that depict this stimulus are perceptually bistable: the apparent rotation direction of the cube reverses spontaneously during continuous viewing. Short movies (1–2 s) are stable but may reverse from one showing to the next. Binocular disparity can be added to make the rightward-moving parts of the image stereo-scopically either closer or farther than leftward moving parts, to bias the apparent rotation direction (e.g., Bradley, Chang, & Andersen, 1998; Dodd, Krug, Cumming, & Parker, 2001).

Stimulus position within the visual field (the POSN cue) is a signal that can be recruited as a cue for a Necker cube's apparent rotation direction if shown in repeated pairing with binocular disparity or with occlusion and binocular disparity. Rotation direction, disparity, and POSN correspond to nodes **A**, **T**, and **N**, respectively, in Figure 2. After training, POSN is effective both in monocular test displays that do not contain binocular disparity signals (Haijiang et al., 2006) and in binocular test displays that do contain binocular disparity signals (Backus & Haijiang, 2007). However, those studies did not compare the effect of POSN in the two types of display for a given trainee.

In the previous studies, the POSN cue had different effectiveness in different trainees, so we predicted not only that the effect of POSN would be the same for a given trainee in the two types of display but also that this equality of effectiveness should hold across a range of effect sizes seen across the trainees.

Methods

Trainees were six undergraduates at the University of Pennsylvania who passed a screening test for stereoacuity (Haijiang et al., 2006). Data were also collected from two additional trainees who passed the stereoacuity screening test but who reported seeing fewer than 80% of training trials to have the rotation direction specified by disparity; their data were not analyzed further.

Anaglyph displays showed brief (2 s) movies of a spotted Necker cube that rotated about a vertical axis (Backus & Haijiang, 2007). The cube subtended 14 deg at the 200 cm viewing distance and was centered 33 cm above or below a fixation mark at the center of the rear projection display screen. Trainees received no feedback on either training trials or test trials. Training trials contained natural binocular disparity for the simulated object (27 arcmin front-to-back disparity) and an occlusion cue (a vertical pole as in Haijiang et al., 2006). These cues controlled apparent rotation direction on training trials. On each trial the trainee pressed a button to indicate apparent rotation direction. Response mapping (i.e., button assignment) was made contingent on a dot that moved either left or right (randomly chosen for that trial) through the fixation mark, to discourage trainees from assigning buttons to the POSN cue *per se*.

Each trainee ran in two sessions each consisting of 480 trials. Training and test trials were presented in alternation. Of the 240 training trials, 120 were in the TOP position (above the fixation mark, Figure 3) and 120 were in the BOTTOM position (below the fixation mark). Half of the trainees were trained on right-hand rotation in the TOP position and the other half were trained on right-hand rotation in the BOTTOM position (left-hand rotation was in the other position). Of the 240 test trials, 80 were monocular (to the right eye only) and 160 were binocular. Of the monocular test trials, 40 were TOP and 40 were BOTTOM. Of the binocular test trials, 80 were TOP and 80 were BOTTOM. Binocular test trials contained scaled binocular disparities that were −4/6, −1/6, +1/6, or +4/6 of the normal disparity signal for a right-rotating

cube (thus, each of the two sessions included 20 trials at each of 4 disparities and 2 POSNs). Test trials of different types were presented in a counterbalanced pseudo-random order.

Additional details of methods are as described in Backus and Haijiang (2007).

Results and discussion

To model apparent rotation direction we took into account four factors: a prior, noise, a trusted cue (binocular disparity), and the newly recruited POSN cue. The perceptual decision on each trial (A_0 = leftward rotation, A_1 = rightward rotation) depended on all four factors. In addition, one would expect the two POSN cues' moments to grow in magnitude during the session. It can be shown that d^* computed for the entire session (by pooling responses on test trials) underestimates the mean value of the cues' difference in moments. However, computing d^* on the entire session is sufficient to test the hypothesis and in any case the experiment did not generate enough data to examine the time course of learning in detail.

Figure 4 shows data from 320 binocular test trials and 160 monocular test trials collected in two sessions from each of six trainees (S1–S6). Percent of trials with rightward apparent rotation is the ordinate. The ordinate therefore estimates the fraction of trials on which the hypothetical decision variable exceeded criterion, i.e., $\Pr(A_1)$ which is to say $\Pr(\hat{\theta} > 0.5)$ for $\hat{\theta}$ in Equation 2 or, equivalently, $\Pr(\mathbf{M} > 0)$ for $\mathbf{M}$ in Equation 4.

Disparity (for the binocular stimuli) is the abscissa. Response at the TOP position is plotted as blue circles (binocular) or a light blue line segment (monocular). Percent rightward at the BOTTOM position is plotted as red stars (binocular) or an orange line segment (monocular). Both disparity and POSN were effective as cues for all trainees and there were significant individual differences in both cues' effectiveness.

In this experiment there was no way to separately estimate θ_i (or ρ_i) and n_i so we used the method of moments instantiated by Equation 4. We assumed that for a given trainee the prior (m_0) was the same on all trials, and that the noise (m_{-1}) was drawn from the standard normal distribution on each trial, now written in bold ($\boldsymbol{m}_{-1}$) to show that it is a random variable. $\boldsymbol{m}_{-1}$ was assumed to be the same for monocular and binocular stimuli.[3]

The terms m_1 and m_2 correspond to the POSN and disparity cues, respectively. For each trainee m_1 takes on two values, $\pm a$, that we can write as ax_1 where x_1 is −1 or +1 and a is a model parameter chosen to maximize the likelihood of the data. The moment of the POSN cue is a and we will plot the effect of the cue as $2a$ so that its size will correspond to the full effect of the POSN cue as measured in z-score units.

That leaves four fixed levels of m_2, which we will assume to be linearly related to disparity: $m_2 = bx_2$ where x_2 is disparity (−4, −1, 1, 4) and b is chosen to maximize the likelihood of the data. The model is now described by

[3]This assumption is an approximation. Whereas the POSN cue was presumably measured without error, so that only lack of trust in POSN as a cue limited its effect on perceptual decisions, error in measuring the disparity cue may have contributed to variability in the perceptual decision. Such noise would contribute to other sources of noise represented by $\boldsymbol{m}_{-1}$ in the simple model we used to analyze the data and would tend to reduce the POSN cue's moment as measured by the model. If disparity measurement noise is significant relative to other sources of noise and if this measurement noise varies in amplitude with the value of disparity itself, then a 2nd noise term that depends on the disparity value would make the model more accurate. Our data are insufficient to test for this. In using the MBE model, progress in understanding perceptual decisions must necessarily proceed through successive explanation of the noise term. The unexplained noise must grow smaller as the perceptual decision comes to be better understood and variance accounted for by specific terms outside the noise. Individual differences in the measurement noise for a given trusted cue are to be expected and could conceivably be exploited to better understand the perceptual decision process.

$$\mathbf{M}=m_{-1}+m_0+ax_1+bx_2, \quad (5)$$

with $\boldsymbol{m}_{-1}$ distributed as standard normal, so $\Pr(A_1)$ can be modeled using probit analysis (Finney, 1971).[4] To the extent that the data in Figure 4 are fit by cumulative Gaussians, we can suppose that the (psychological) moment for disparity as a cue was indeed proportional to the (physical) disparity in the stimulus and that the noise was normally distributed. To compute 2*a* on monocular test trials, *b* was set to 0.

Fits were better when the cumulative Gaussians in Figure 4 were forced to asymptote not at 0% and 100% but rather at ε/2 and 100 − ε/2, where ε was the percentage lapse rate, estimated as twice the error rate for training trials. Lapse rates were 0.0%, 1.3%, 0.4%, 4.6%, 1.7%, and 1.3% for S1 through S6, respectively. This correction prevents systematic bias in the computation of the regression coefficients (Swanson & Birch, 1992;Wichmann & Hill, 2001). None of our basic conclusions were affected by modeling the lapse rate.[5]

Finally, Figure 5 plots the effect of POSN in binocular stimuli as a function of its effect in monocular stimuli for each of the six trainees. The diagonal line with unit slope is the prediction for a system that interprets the new POSN cue as though it is conditionally independent of the long-trusted binocular disparity cue. Boxes show 50% confidence intervals and bars show 95% confidence intervals. Two individuals (S1 and S6) deviated from the prediction with statistical significance ($p < 0.05$) assuming bivariate normal error but deviations from this line are not systematic and the fit appears to be quite good, considered at the population level. The 95% confidence interval for the slope of the line that passes through these six trainees' data points is (0.70, 1.70) as calculated from the 2.5 and 97.5 percentiles in a distribution of 1000 estimates (resulting from linear fits to six data points, not constrained to pass through the origin, with each data point in a given bootstrap sample being determined by resampling the data from a given trainee).

That the data fall along the prediction line suggests that cue recruitment, as a form of perceptual learning, generally occurs under the naive Bayes assumption. That there were unsystematic but individually reliable deviations from the prediction shows that individual visual systems do not implement this strategy exactly.

It should be noted that test trials were intermixed with training trials in this experiment. The learning was therefore a mixture of short-term and long-term effects (Backus & Haijiang, 2007; Haijiang et al., 2006) and the data do not distinguish between learning at one or another time scale.

Discussion

An important reason for developing MBE model was to make perceptually bistable stimuli theoretically suitable for studies of cue combination and cue recruitment. MBE justifies the use of probit analysis to quantify cue effects. Dosher et al. (1986) previously used probit analysis to quantify cue effects so it is interesting to consider why their methodology was not

[4]Let *Y* be an indicator variable such that *Y* is 1 for a given trial if and only if $\mathbf{M} > 0$ on that trial. Then $\Pr(A_1) = \Pr(Y = 1|m_0, m_1, m_2) = \Phi(m_0 + ax_1 + bx_2)$ where Φ is the cumulative distribution function of the standard normal distribution.

[5]The straightforward probit model was fitted to the data in Figure 3 using Matlab's general linear model fitting function (glmfit for binomially distributed data with a probit linking function). The lapse rate model was fitted by maximizing the likelihood of the data for a pair of parallel cumulative Gaussian probability functions, with the moment for the POSN cue being computed as the horizontal distance between the curves (in units of binocular disparity) divided by their common standard deviation (also in units of binocular disparity). When the lapse rate is not modeled, the psychometric curves in Figure 3 are less steep and the fitted moment for POSN is reduced in size.

widely adopted. First, although they pointed out that the additivity of "evidence" terms was consistent with existing theories of perceptual decision making (and energy landscape theories in particular), they did not provide an interpretation of the belief terms beyond describing them as "strength of evidence." It is important to understand why a measured cue provides "evidence." To assess evidence correctly the system must measure the cue (which will be done with noise), interpret the measured cue (according to an estimate of the cue's ecological validity), and estimate how well it has done these two things. MBE defines terms that relate these theoretically important quantities to the terms in a probit analysis.

Second, most studies of bistability have used prolonged viewing to encourage perceptual rivalry. This history undoubtedly contributes to a lingering suspicion that dichotomous percepts are somehow exceptional. In fact, however, rivalry (or its time course) and MBE have little to say about one another. Short-duration stimuli are more representative of ordinary perception, and more amenable to analysis with MBE, than are long-duration stimuli. At any rate, we believe that students of perception need no longer consider bistable stimuli to be exceptional, and that it is appropriate to measure a cue's effect on a dichotomous perceptual decision as a moment or subjective reliability using probit analysis.

Weak fusion, naive Bayes classifiers, conditional independence, and configural learning

The models described here can be seen as assuming a "weak fusion" theory of cue combination (Clark & Yuille, 1990; Johnston, Cumming, & Parker, 1993; Landy, Maloney, Johnston, & Young, 1995; Young, Landy, & Maloney, 1993) because the perceptual system is assumed to use the cues as independent sources of information to choose between A_0 and A_1. The likelihood for each cue is assumed not to depend on the values of other cues, but only upon some property of the world; the cues are *conditionally independent* so their dependence on the same scene parameter fully explains their covariance. This method of combining information to categorize a stimulus is called a "naive Bayes classifier" in the machine learning literature (Lewis, 1998).

Naive Bayes classifiers have been remarkably successful in practical applications (e.g., Kohavi, 1996), in part because they work well even when conditional independence is violated (Domingos & Pazzani, 1996; Hand & Yu, 2001; Zhang, 2004). Of particular interest here is the possibility that a newly recruited cue is assumed by the system to be conditionally independent of any trusted cues that were used to train it. This assumption, that cues are conditionally independent until proven otherwise, is reminiscent of an experimental finding in the animal learning literature. In "configural learning" experiments, an animal is trained under conditions in which Cue 1 predicts reward and Cue 2 predicts reward, but Cues 1 and 2 together predict no reward. Initially, the animal responds whenever either cue is present, and inappropriately, it responds most vigorously when both cues are present together. With time the animal learns that the conditional independence assumption was incorrect and ceases to respond when both cues are present together (Bellingham, Gillette-Bellingham, & Kehoe, 1985; Woodbury, 1943).

The models presented in this paper provide a framework for testing whether or not the perceptual system assumed conditional independence when recruiting a new cue. Future experiments may provide more precise tests of this hypothesis; if necessary it would be straightforward to add an interaction term to the model equations, to measure the contribution of a new cue and a trusted cue in combination.

MBE and Bayes rule

MBE was developed to quantify the effectiveness of a cue, not as a model for the learning that occurs during cue recruitment. However it naturally suggests a model for this learning in which

each training trial is considered additional evidence that increases confidence in the system's knowledge about the value of θ_i in accordance with Bayes rule. In other words, each trial of the experiment might be like observing one new trial in a binomial experiment. In that case, the posterior *beta* distribution for θ_i (or ρ_i) after each trial becomes the prior for θ_i (or ρ_i) on the next trial, with n_i increasing by 1 on each trial.

This simple implementation of Bayes rule predicts that during a cue recruitment experiment, the effectiveness of a new cue will increase monotonically. This prediction has not been borne out in practice. For some trainees the POSN cue is more effective on average during the first half of a session than the second half (unpublished data; see also the bottom two panels of Figure 3; Backus & Haijiang). Deviations of this sort reflect unexpected dynamics in the mechanisms that implement the learning and will require further study.

In addition, MBE would require at least one additional parameter to implement Bayes rule. A fact about the model is that training trials under these assumptions cause unit increments or decrements in a cue's moment, no matter how many training trials have occurred previously, as the factor n_i in the moment cancels the denominator in the estimate of θ_i (Equation 3). Since $n_{-1} = 1$ by construction, an additional model parameter must be introduced to capture relative scale between the noise and cue terms. Otherwise the model would specify a fixed (and very high) learning rate: after only 4 training trials, **M** would be distributed as $N(4, 1)$ and a perceptual choice of form A_1 on the next test trial would be guaranteed.

A scaling term between the noise and cues is theoretically justified by the need to explain different learning rates for different new cues (Haijiang et al., 2006; Michel & Jacobs, 2007), but the problem of nonmonotonic learning remains. It is clearly important to understand the dynamics of the learning during cue recruitment, but the simple idea that Bayes rule is implemented in a straightforward manner at the time scale of a single experimental trial appears to be too simple.

The general prior

The word "prior" is used two ways in Bayesian statistics. The first meaning is the prior probability distribution for a particular variable, before one has measured cues that would modify one's belief about the variable's value. It is in this sense that 0th term in MBE describes "prior probability"—it is the overall probability of the world being in the state that corresponds to A_1, i.e., the system's best guess in the absence of data. In a coherent belief system (Lindley, Tversky, & Brown, 1979; Osherson, Shafir, Krantz, & Smith, 1997), this prior distribution is equal to the marginal distribution for that variable within the *general prior*, which is the second meaning of the term. The general prior is a high dimensional space of joint probability density in which each possible measured signal and each parameter describing a state of the world has its own dimension. Statistical relationships between different signals, or between different states of the world, or between signals and states of the world, can be represented by the distribution of probability density within the general prior.

MBE is embedded within this Bayesian framework in that a perceptually bistable stimulus is one for which the set of signal measurements restrict (or condition) the general prior to just two main maxima of local mass (Brainard & Freeman, 1997), both of which are attractive to a representational apparatus whose job it is to guess the most likely state of the world. We must further suppose that the representational apparatus will choose just one of the two maxima at any given time. Cues (such as binocular disparity) bias the percept by causing one maximum to have more local mass than the other, with decision noise also affecting the final choice.

Within this framework, perceptual learning is the updating of the general prior by mechanisms whose job it is to keep perception accurate. Different types of perceptual learning correspond

to different types of updating. Improved ability discriminate requires the concentration of probability density at the nexus of the relevant cues and a decision variable; this sharpens the weighting function across many cues that participate in construction of the decision variable (Petrov, Dosher, & Lu, 2005). A recalibration is a shift of probability density between regions of the general prior, and cue recruitment is the deposition of probability density for a single cue where previously there was no such concentration.

Probability matching as an alternative decision rule

In MBE, a source of noise is added to the decision variable to account for trial-to-trial variability in the appearance of the stimulus as A_0 or A_1. The decision rule itself is deterministic. Alternatively, one could suppose that the decision rule is probabilistic. Thus we might entertain a model in which the system constructs the posterior for θ as in MBE but without any (−1st) noise term, with the decision rule being to choose A_1 with probability $\hat{\theta}$ [i.e., choose A_1 if a uniformly distributed random variable on [0, 1] is less than $\hat{\theta} = E(\pi(\theta))$].

This strategy is a "probability matching" strategy (Herrnstein, 2000). Probability matching was previously adopted by Mamassian and Landy (1998, 2001) as a solution to the problem of explaining decision variability in perceptually dichotomous tasks. This solution is appealing for its simplicity, but theoretically we see little justification for it as a near-optimal strategy in perception. We presume the existence of perceptual mechanisms whose job it is to promote alternative interpretations of the sense data, but it seems unlikely that the cost function for perceptual decisions would be exactly the one for which optimal perception requires seeing the world one way versus the other in proportion to the system's internal estimate of probability. Indeed, the probability matching model is awkward if *any* noise in the perceptual decision comes to be understood, because each such discovery reveals that previous model was not correctly describing the construction of $\hat{\theta}$. MBE does not suffer this problem (see Meaning of the noise term section below).

Animals and humans do exhibit probability matching or near-matching in a variety of behavioral situations, even when this strategy does not maximize rewards (Gallistel, 1990). One explanation for probability matching is that organisms choose at any moment to do that which yields the highest reinforcement, with a claim that this so-called "melioration" yields well-adapted behaviors in natural environments (Herrnstein, 2000). Could probability matching operate the same way, and serve the same purpose, in perception as in other behaviors? In the case of a foraging animal that must compete with other animals for food, probability matching is an evolutionarily stable strategy because the animal is more likely to find food by taking into account where other foragers are. But there is no comparable competition for scare resources in the case of perception. Matching has also been invoked as strategy for balancing the exploitation of current knowledge with the need to detect changes in contingency (e.g., Sabes & Jordan, 1996; Vulkan, 2000). It is conceivable that perceptual probability matching is somehow optimal to facilitate the detection of changes in contingency for visual cues, analogous to a detection of change in contingency for food cues, but we do not know of evidence to support this idea.

Contributors to subjective reliability

In the Brunswikian tradition, Stewart and Lusk (1994) distinguish between "reliability of information acquisition" and "reliability of information processing" when an organism estimates a property of the world from a cue. The former describes the loss of information that occurs during measurement of the cue by the system, while the latter describes failure to make appropriate use of the measured cue. This distinction is similar to that between intrinsic noise and efficiency (Pelli & Farell, 1999). Both are distinct from the cue's ecological validity (correlation with the property in the world) prior to measurement by the system.

Ecological validity prior to measurement and reliability of information acquisition (error in cue measurement) are together responsible for spread in the likelihood function of the measured cue. In other words, the pattern of neural activity evoked by a property of the world at a given instant depends probabilistically on both how good the cue is and on how well the cue is measured by the system. The system's ability to infer the property of the world is then further limited by the reliability of additional information processing, i.e., how efficiently the system uses the measured cue during the construction of appearance. Efficient usage has yet additional requirements: that the cue be transmitted faithfully as an input to a perceptual computation, that the computation have the correct functional form, and that information in the cue be properly weighted in comparison to other independent sources of information about the property of the world.

In MBE, all of this process with the exception of cue weighting is realized by the hypothetical estimation of θ_i, or equivalently, estimation of the cue's ecological validity ρ_i. (Recall that for notational simplicity we did not mark the distinction between the true value of ρ_i and the system's estimate of ρ_i. In MBE, as in any Bayesian model, a given act of perception is based on the estimate). The cue's subjective reliability, $m_i = n_i\rho_i$, is thus a confidence-weighted estimate of ecological validity that reflects both the reliability of acquisition and the reliability of processing. Because a real system cannot perfectly know the statistics of the current environment, the statistics governing signal measurement, nor whether it is using signals efficiently, it cannot be certain how to distribute relative trust to its experts in cases where they recommend opposing decisions. Thus we must suppose that the weight n_i is also an estimate, one that may be refined over time as the system discovers according to some learning rule that the *i*th cue is more or less reliable.

Subjective reliability for binary and continuously valued perceptual attributes

Knill (2003), in discussing the construction of a continuously valued perceptual attribute, uses the term "subjective reliability" in a manner that can be made consistent with our use in MBE. As in MBE, Knill used the term to capture the system's actual reliance on a cue, as opposed to the reliance one would expect if the system were properly calibrated to make optimal use of the cue's actual reliability (defined by Backus & Banks, 1999, to be the reciprocal variance across repeated trials in the system's estimates). However, these schemes for characterizing reliability are not directly applicable to dichotomous decisions because they measure reliability in units of reciprocal squared continuous units (e.g., $1/\text{deg}^2$ in the case of perceived surface slant). A formalization that relates subjective reliability for continuous perceptual attributes to that for binary perceptual attributes is thus called for.

A straightforward application of Knill's approach to dichotomous perceptual decisions can be realized if we suppose ρ_i to be binary, with a value of −1 or 1 on each trial. In that case, n_i has the same magnitude as m_i so that a cue's weight becomes equal to its subjective reliability. From this it follows trivially that a cue's relative weight during cue combination is proportional to its subjective reliability. This formulation does not capture the fact that some cues have higher ecological validity than others, but it does allow us to see that a cue's subjective reliability in MBE can be identified with subjective reliability in a model of cue combination for nonbinary perceptual attributes. Indeed, our development of MBE shows why similar theory may be needed to explain why subjective reliability differs from actual reliability in the case of continuously valued perceptual attributes.

In the cue recruitment experiment described here, we measured a newly recruited cue's moment, which was interpreted as subjective reliability. We did not separately measure its estimated ecological validity and weight, so the data do not on their own require MBE. However, the experiment became theoretically meaningful in consequence of MBE. Without

theory we could not have justified the use of probit analysis, and our adoption of the term "subjective reliability" for the new cue's weight would have been arbitrary.

Future cue recruitment experiments should be able to determine whether it is *necessary* to posit separate estimates for ecological validity and weight. MBE predicts that a given subjective reliability should be attainable using either (1) many training trials across which the correlation between the new and trusted cues is low (leading to small ρ_i and large n_i) or (2) fewer training trials with high correlation (larger ρ_i and smaller n_i). If additional training trials are differently effective in these two situations (we predict that they will be more effective in the second case), then a two-parameter model such as MBE will be needed to describe the two states of learning, respectively.

Meaning of the noise term

The noise term in MBE represents what the experimenter does not know about how the perceptual choice was made on a given type of trial. Thus, additional predictor variables can reduce the noise in the model, without of course changing anything about the perceptual system. Each of the curves in Figure 4 might decompose into a pair of steeper, horizontally separated curves, if we add some other predictor to the model. Such predictors might include the 3D starting orientation of the cube in the movie, or the system's decision on the previous trial. If the system relies on some cue that varies but that the experimenter has not modeled, then reliance on the modeled cues (as measured for example by their moments) will be lower than if the unmodeled cue is included.

Similarly, after training the system to use POSN as a cue we might prevent the experimenter from knowing what the value of POSN was on each trial. In that case the noise term would grow, and the moment for the disparity cue (which is measured in normal equivalent deviations) would be relatively less effective in controlling the appearance of the stimulus than is actually the case (and indeed, less effective than it was before position was conditioned to act as a cue). Of course, MBE also allows us to test whether the effectiveness of a long-trusted cue changes after a new cue is recruited.

An alternative approach is to characterize belief for each cue in absolute terms, using log-odds (see Appendix A). This approach would be preferable if it were possible: the subjective reliabilities of a cue and of the noise could then be given absolute values, rather than being measured relative to one another. We do not see how to implement this approach except as a computer model, however, because we do not yet have sufficient understanding of the noise processes that contribute to the perceptual decision. As a result, we are not yet in a position to determine whether a small cue moment results from lack of trust in the cue or from high decision noise.

Conclusion

MBE is a theory that provides a rationale for expecting that the effects of independent cues on the appearance of a perceptually bistable stimulus should be additive when expressed as *z*-scores (i.e., in units of normal equivalent deviations or NEDs). An assumption of conditional independence was exploited to develop the modeling equations, but it is straightforward to test for interactions using standard techniques of probit analysis. The construction of binary-valued perceptual attributes from perceptually bistable stimuli need not be considered exceptional insofar as it can now be handled within a Bayesian framework. It therefore seems acceptable to exploit perceptually bistable stimuli to study the mechanisms by which perceptual appearance is constructed. Data from a cue recruitment experiment were consistent with a hypothesis that by default, the visual system makes the assumption that a newly discovered

cue is conditionally independent of the long-trusted cues from which the new cue's meaning was inferred.

Appendix A

MBE vs. log-odds

MBE leads to probit analysis to quantify each cue's effectiveness. An alternative to probit analysis is a log-odds approach (e.g., Mammassian & Landy, 1998,2001). For some hypothesis tests it does not matter whether one uses probit analysis or log-odds to fit the data: data that are fit by one model are often very well fit by the other, and differences between conditions that are statistically significant using one model are likely to be significant using the other. As theories to describe perception, however, these approaches differ. In MBE, each cue contributes a subjective reliability that can be decomposed into an ecological validity and a weight, whereas in the log-odds approach each cue contributes a single number to overall belief. When using log-odds, the assumption that the belief terms for different cues are additive implies an assumption that the cues are conditionally independent and vice versa (because the likelihoods for different cues are multiplied). In MBE one could build a model in which belief terms were additive without committing to the cues' conditional independence.

Here we consider two ways to implement a log-odds model for combining expert opinions in the service of a dichotomous perceptual decision. These two approaches lead to identical modeling equations, and in fact this equation is similar to Equation 4 for the grand moment in MBE. The meaning (theoretical interpretation) of the equation is different in the three cases, however.

Independent experts that report Pr($A_1|x_i$)

The first log-odds approach is to assume that the *i*th expert accurately reports the likelihood $\Pr(x_i|A_1)$ where x_i is the measured cue used by that expert. Assuming the expert also knows $\Pr(x_i|A_0)$, $\Pr(A_1)$, and $\Pr(A_0) = 1 - \Pr(A_1)$, the expert could, equivalently, report a probability, $\theta_i = \Pr(A_1|x_i)$, as in the MBE model. However, this number would be much more powerful than θ_i when it appeared in Equation 2: it is no longer subject to a separate confidence evaluation in the form of n_i, but rather estimates a true probability that we must take at face value. There is no way within this model to account for the possibility that one expert's estimate of θ is of better quality than another's. However, if we trust our N experts to report $\{\theta_i\}$ accurately, we can calculate a subjective probability for the state of the world because $\hat{\theta}$ will approximate θ.

The system's estimated log-odds (LO) in favor of form A_1, in a situation where A_0 or A_1 are the only viable alternatives, is

$$LO = \ln \frac{\Pr(A_1|\mathrm{x})}{\Pr(A_0|\mathrm{x})} = \ln \frac{\theta}{1-\theta}, \tag{A1}$$

where $\theta = \Pr(A_1|\mathbf{x})$ and $\mathbf{x} = (x_1, x_2, \ldots, x_N)$ is a set of measurements made by the system that are relevant to determining θ (i.e., the set of measurements that depend on whether the binary state of the world corresponds to A_0 or A_1, respectively).

Using Bayes theorem Equation A1 can be rewritten as:

$$LO = \ln \frac{\Pr(A_1)\Pr(\mathrm{x}|A_1)}{\Pr(A_0)\Pr(\mathrm{x}|A_0)} = \ln \frac{\theta_0 \Pr(\mathrm{x}|A_1)}{(1-\theta_0)\Pr(\mathrm{x}|A_0)}, \tag{A2}$$

or

$$LO=\ln\frac{\theta_0}{(1-\theta_0)}+\ln\frac{\Pr(\mathrm{x}|A_1)}{\Pr(\mathrm{x}|A_0)}=\ln K_0+\ln K, \tag{A3}$$

where $\theta_0 = \Pr(A_1)$ is the prior probability in favor of form A_1 and $\ln K_0$ is the log-odds for prior belief in favor of A_1. K is called the Bayes factor (or likelihood ratio) and $\ln K$ is the *weight of evidence* in favor of A_1 (Good, 1950). Like **M**, LO is positive when $\hat{\theta} > 0.5$, has range $[-\infty, +\infty]$, and decomposes the system's belief into prior belief and an update of that belief based on evidence. There is no separate noise term, however.

K can be decomposed into meaningful terms that correspond to the evidence provided by separate cues, if we assume conditional independence. Suppose therefore that $\mathbf{x} = (x_1, x_2, \ldots, x_N)$ with x_i being the cue measurement upon which the ith expert depends, $i = 1, 2, \ldots, N$. The values of $x_1, x_2, \ldots, x_N$ thus depend on the state of the world, but not on each other. Exploiting the conditional independence we can write

$$\Pr(\mathrm{x}|A_k)=\prod_{i=1}^{N}\Pr(x_i|A_k),\ k=0 \text{ or } 1, \tag{A4}$$

so that

$$LO=\ln K_0+\sum_{i=1}^{N}\ln\frac{\Pr(x_i|A_1)}{\Pr(x_i|A_0)}=\sum_{i=0}^{N}\ln L_i, \tag{A5}$$

where L_i is the likelihood ratio for the ith cue (with $L_0 = K_0$). Equation A5 expresses the decision variable LO in terms of a prior belief about θ (instantiated by $\ln L_0$) and updates to the prior based on the likelihood ratios (odds) of the measurements $\{x_i\}$.

For comparison with Equation 2, we can recast Equation A5 in terms of the experts' probability reports, $\theta_i = \Pr(A_1|x_i)$. Applying Bayes theorem again to Equation A4 we obtain

$$\begin{aligned}\ln K &= \ln\prod_{i=1}^{N}\frac{\Pr(A_1|x_i)}{\Pr(A_0|x_i)}+\ln\left[\frac{\Pr(A_0)}{\Pr(A_1)}\right]^N \\ &= \sum_{i=1}^{N}\ln\left(\frac{\theta_i}{1-\theta_i}\right)-N\ln\left(\frac{\theta_0}{1-\theta_0}\right),\end{aligned} \tag{A6}$$

which combines with Equation A3 to yield

$$LO=(1-N)\ln K_0+\sum_{i=1}^{N}\ln K_i, \tag{A7}$$

where $K_i = \theta_i / (1 - \theta_i)$ is the odds for a decision based on ith cue alone. This formulation shows how $\{\theta_i\}$ can be combined using a log-odds approach when each expert provides an independent estimate of θ. The decision maker takes these estimates at face value, which puts all onus of estimating confidence in the expert's skill onto the expert. In MBE, this role is

played by the cue's weight and the decision maker need not accept the expert's estimate of weight at face value.

Note also that in Equation A7 the influence of the prior, shown by the first term, depends on N, the number of cues. This result may seem counterintuitive, but recall that each expert had to separately take the prior into account. The first term therefore factors out this correlated component from the experts' respective contributions.

Experts reporting Pr(A_1 is the best choice|x_i)

The second log-odds approach allows each expert to report not Pr($A_1|x_i$) but rather Pr(A_1 is the best choice|x_i), in other words, the probability that choosing A_1 maximizes percent correct. This gives the expert a chance to hedge: we do not ask that experts report Pr($A_1|x_i$), but only their subjective probability for the proposition that choosing A_1 is the right thing to do to maximize percent correct. One can defend this approach by assuming, as before, that Pr($A_1|x_i$) exists as a single number, but that each independent expert knows this probability only up to a probability distribution, or more specifically, only up to the area of the probability distribution to the right of 0.5.

Equations A1–A7 can be re-used for this second approach by replacing $\theta = \Pr(A_1|\mathbf{x})$ with Pr(A_1 is the best choice|$\mathbf{x}$), and $\theta_i = \Pr(A_1|x_i)$ with Pr(A_1 is the best choice|x_i). Imprecise estimates of L_i or K_i (in their new meanings) can also again be used according to Equation A5 or A7 to predict whether A_1 is the best choice.

Comparison of the three approaches

As process models, both *LO* approaches give all power over the final decision to the experts, with no role for a decision maker who looks at their reports. The second of these *LO* approaches may be theoretically more plausible as a model for perceptual decisions, as it more naturally allows each expert to raise or lower the subjective probability it reports (depending on an estimate of reliability that is internal to the expert). These two log-odds approaches are alternatives to Equation 4 to express in a model how cues combine in a decision about whether A_0 or A_1 represents the state of the world.

From the experimenter's point of view, **M** in Equation 4 and *LO* in Equation A5 or A7 are indistinguishable, provided we assume a fixed criterion and add an extra term to the equation for *LO* to represent a source of decision noise (m_{-1}, or $\ln L_{-1} = \ln K_{-1}$). In that case probit analysis could be used to measure the log-odds for each cue—in units of normal-equivalent-deviations-of-log-odds-noise. All three equations are simple sums in which each expert contributes one term. By adding and subtracting cues from a stimulus, and by changing their levels, the experimenter can measure the effect of a given cue according to its ability to influence the perceptual decision using any of these models. It is at present unclear which of the three approaches best describes perception.

It seems prudent to have in the back of one's mind that the system's confidence in the *i*th expert should be a separate parameter. In MBE we represented this confidence using the weight n_i in Equation 2. The *i*th expert could help decide the value of n_i on a given trial, but at least some part of the determination of n_i should be made at the system level where the expert's report can be compared to the reports of the other experts, and updated over time. MBE quite naturally explains each belief term as a combination of confidence and ecological validity and one could presumably decompose a log-odds belief term into corresponding components as well. One such effort is described by Vermunt and Magidson (2003). Additional methods to treat uncertainty in a log-odds formulation can be found in Savage (1972) and Lindley et al. (1979).

In any case we prefer to add decision noise as a separate (−1st) noise term in using *LO*. Otherwise the decision noise will be distributed among the various belief terms for the cues, so the model would not contain terms for the log-odds contributions of the cues to the perceptual decision, but only for their log-odds contributions to the final response. Log-odds contribution to the response is a meaningful quantity, but hiding the decision noise is ultimately not helpful when the ultimate goal is to understand the factors that contribute to a dichotomous perceptual decision.

Acknowledgments

This research was supported by NIH Grant EY-013988, the Human Frontier Research Program, and NSF Grant BCS-0810944. The author thanks Qi Haijiang for writing computer code to run the experiment, Sam Cohn and Michael Filetti for running subjects, Dean Foster for discussion of the statistical learning literature, and Michael Landy and an anonymous reviewer for helpful comments on the manuscript.

References

Attneave F. Multistability in perception. Scientific American 1971;225:63–71. [PubMed: 5116412]

Backus BT, Banks MS. Estimator reliability and distance scaling in stereoscopic slant perception. Perception 1999;28:217–242. [PubMed: 10615462]

Backus BT, Haijiang Q. Competition between newly recruited and pre-existing visual cues during the construction of visual appearance. Vision Research 2007;47:919–924. [PubMed: 17303207]

Barlow H. Conditions for versatile learning, Helmholtz's unconscious inference, and the task of perception. Vision Research 1990;30:1561–1571. [PubMed: 2288075]

Bellingham WP, Gillette-Bellingham K, Kehoe EJ. Summation and configuration in patterning schedules with the rat and rabbit. Animal Learning & Behavior 1985;13:152–164.

Bradley DC, Chang GC, Andersen RA. Encoding of three-dimensional structure-from-motion by primate area MT neurons. Nature 1998;392:714–717. [PubMed: 9565031]

Brainard DH, Freeman WT. Bayesian color constancy. Journal of the Optical Society of America A, Optics, Image Science, and Vision 1997;14:1393–1411.

Brunswik, E. Perception and the representative design of psychological experiments. Berkeley, CA: University of California Press; 1956.

Clark, JJ.; Yuille, AL. Data fusion for sensory information processing systems. Boston: Kluwer; 1990.

Clemen RT, Winkler RL. Combining probability distributions from experts in risk analysis. Risk Analysis 1999;19:187–203.

Coombs, CH.; Dawes, RM.; Tversky, A. Mathematical psychology: An elementary introduction. Englewood Cliffs, NJ: Prentice-Hall; 1970. The theory of signal detectability (Ch 6); p. 165-201.

Dietterich, TG. Machine learning for sequential data: A review. In: Caelli, T., editor. Structural, syntactic, and statistical pattern recognition; lecture notes in computer science. Vol. vol. 2396. London: Springer-Verlag; 2002. p. 15-30.

Dodd JV, Krug K, Cumming BG, Parker AJ. Perceptually bistable three-dimensional figures evoke high choice probabilities in cortical area MT. Journal of Neuroscience 2001;21:4809–4821. [PubMed: 11425908]

Domingos, P.; Pazzani, M. Beyond independence: Conditions for the optimality of the simple bayesian classifier; Thirteenth International Conference on Machine Learning (ICML); 1996.

Dosher BA, Sperling G, Wurst SA. Tradeoffs between stereopsis and proximity luminance covariance as determinants of perceived 3D structure. Vision Research 1986;26:973–990. [PubMed: 3750879]

Fechner, GT. Elemente der psychophysik. Leipzig, Germany: Breitkopf und Härtel; 1860.

Feldman J. Bayesian contour integration. Perception & Psychophysics 2001;63:1171–1182. [PubMed: 11766942]

Feldman J, Tremoulet PD. Individuation of visual objects over time. Cognition 2006;99:131–165. [PubMed: 16545625]

Fine I, Jacobs RA. Comparing perceptual learning tasks: A review. Journal of Vision 2002;2(2):5, 190–203.http://journalofvision.org/2/2/5/

Finney, DJ. Probit analysis. Cambridge, UK: Cambridge University Press; 1971.

Foster D, Vohra R. Regret in the on–line decision problem. Games and Economic Behavior 1999;29:7–36.

Gaddum, JH. Medical Research Council Special Report Series (No. 183). London: H.M. Stationary Office; 1933. Reports on biological standards. III. Methods of biological assay depending on a quantal response.

Gallistel, CR. The organization of learning. Cambridge, MA: MIT Press; 1990.

Geisler WS. Sequential ideal-observer analysis of visual discriminations. Psychological Review 1989;96:267–314. [PubMed: 2652171]

Geisler WS, Kersten D. Illusions, perception and Bayes. Nature Neuroscience 2002;5:508–510.

Geisler WS, Perry JS, Super BJ, Gallogly DP. Edge co-occurrence in natural images predicts contour grouping performance. Vision Research 2001;41:711–724. [PubMed: 11248261]

Gelman, A.; Carlin, JB.; Stern, HS.; Rubin, DB. Bayesian data analysis. London: Chapman & Hall; 1995.

Good, IJ. Probability and the weighing of evidence. London: Charles Griffin; 1950.

Green, DM.; Swets, JA. Signal detection theory and psychophysics. New York: Krieger; 1966.

Haijiang Q, Saunders JA, Stone RW, Backus BT. Demonstration of cue recruitment: Change in visual appearance by means of Pavlovian conditioning. Proceedings of the National Academy of Sciences of the United States of America 2006;103:483–486. [PubMed: 16387858]

Hand DJ, Yu K. Idiot's Bayes—Not so stupid after all? International Statistical Review 2001;69:385–398.

Hebb, DO. Organization of behavior. New York: Wiley; 1949.

Helmholtz, Hv. *Treatise on physiological optics* (vol III, J.P.C. Southall, transl. from German). New York: Optical Society of America; 19101925.

Herrnstein, RJ. The Matching Law: Papers in Psychology and Economics. Rachlin, H.; Laibson, DI., editors. Cambridge, MA: Harvard University Press; 2000.

Hillis JM, Watt SJ, Landy MS, Banks MS. Slant from texture and disparity cues: Optimal cue combination. Journal of Vision 2004;4(12):1, 967–992. [PubMed: 14995894] http://journalofvision.org/4/12/1/

Ho YX, Landy MS, Maloney LT. Conjoint measurement of gloss and surface texture. Psychological Science 2008;19:196–204. [PubMed: 18271869]

Ho YX, Maloney LT, Landy MS. The effect of viewpoint on perceived visual roughness. Journal of Vision 2007;7(1):1, 1–16. [PubMed: 17461669]http://journalofvision.org/7/1/1/

Hochberg JE, Krantz D. Brunswik and Bayes: A review of "The Essential Brunswik: Beginnings, Explications, Applications" by Kenneth R. Hammond and Thomas R. Stewart (2001). Contemporary Psychology: APA Review of Books 2004;49:785–787.

Hogervorst MA, Eagle RA. Biases in three-dimensional structure-from-motion arise from noise in the early visual system. Proceedings of the Royal Society B: Biological Sciences 1998;265:1587–1593.

Hummel RA, Landy MS. A statistical viewpoint on the theory of evidence. IEEE Transactions on Pattern Analysis and Machine Intelligence 1988;10:235–247.

Johnston EB, Cumming BG, Parker AJ. Integration of depth modules: Stereopsis and texture. Vision Research 1993;33:813–826. [PubMed: 8351852]

Kersten D, Mamassian P, Yuille A. Object perception as Bayesian inference. Annual Review of Psychology 2004;55:271–304.

Knill, D.; Richards, W. Perception as Bayesian inference. London: Cambridge University Press; 1996.

Knill DC. Mixture models and the probabilistic structure of depth cues. Vision Research 2003;43:831–854. [PubMed: 12639607]

Kohavi, RI. Scaling up the accuracy of naive-Bayes classifiers: A decision-tree hybrid; Paper presented at the Second International Conference on Knowledge Discovery and Data Mining; 1996.

Landy MS, Maloney LT, Johnston EB, Young M. Measurement and modeling of depth cue combination: In defense of weak fusion. Vision Research 1995;35:389–412. [PubMed: 7892735]

Leopold DA, Logothetis NK. Multistable phenomena: Changing views in perception. Trends in Cognitive Sciences 1999;3:254–264. [PubMed: 10377540]

Lewis, DD. Naive Bayes at forty: The independence assumption in information retrieval; Paper presented at the Tenth European Conference on Machine Learning; Berlin. 1998.

Lindley DV, Tversky A, Brown RV. On the reconciliation of probability assessments. Journal of the Royal Statistical Society Series A: General 1979;142:146–180.

Long GM, Toppino TC, Kostenbauder JF. As the cube turns: Evidence for two processes in the perception of a dynamic reversible figure. Perception & Psychophysics 1983;34:29–38. [PubMed: 6634356]

Mamassian P, Landy MS. Observer biases in the 3D interpretation of line drawings. Vision Research 1998;38:2817–2832. [PubMed: 9775328]

Mamassian P, Landy MS. Interaction of visual prior constraints. Vision Research 2001;41:2653–2668. [PubMed: 11520511]

Michel MM, Jacobs RA. Parameter learning but not structure learning: A Bayesian network model of constraints on early perceptual learning. Journal of Vision 2007;7(1):4, 1–18. [PubMed: 17461672] http://journalofvision.org/7/1/4/

Morris PA. An axiomatic approach to expert resolution. Management Science 1983;29:24–32.

Necker LA. Observations on some remarkable optical phenomena seen in Switzerland; and on an optical phenomenon which occurs on viewing a figure of a crystal or geometrical solid. London and Edinburgh Philosophical Magazine and Journal of Science 1832;1:329–337.

Osherson D, Shafir E, Krantz DH, Smith EE. Probability bootstrapping: Improving prediction by fitting extensional models to knowledgeable but incoherent probability judgments. Organizational Behavior and Human Decision Processes 1997;69:1–8.

Pelli DG, Farell B. Why use noise? Journal of the Optical Society of America A, Optics, Image Science, and Vision 1999;16:647–653.

Petrov AA, Dosher BA, Lu ZL. The dynamics of perceptual learning: An incremental reweighting model. Psychological Review 2005;112:715–743. [PubMed: 16262466]

Sabes, PN.; Jordan, MI. Reinforcement learning by probability matching; Paper presented at the Advances in Neural Information Processing Systems; Cambridge, MA. 1996.

Savage, LJ. Foundations of statistics. New York: Courier Dover Publications; 1972.

Shafer, G. A mathematical theory of evidence. Princeton, NJ: Princeton University Press; 1976.

Shafer G. Constructive probability. Synthese 1981;48:1–60.

Smith JE, von Winterfeldt D. Decision analysis in management science. Management Science 2004;50:561–574.

Stewart TR, Lusk CM. 7 components of judgmental forecasting skill—Implications for research and the improvement of forecasts. Journal of Forecasting 1994;13:579–599.

Swanson WH, Birch EE. Extracting thresholds from noisy psychophysical data. Perception & Psychophysics 1992;51:409–422. [PubMed: 1594431]

van Ee R, Adams WJ, Mamassian P. Bayesian modeling of cue interaction: Bistability in stereoscopic slant perception. Journal of the Optical Society of America A, Optics, Image Science, and Vision 2003;20:1398–1406.

Vermunt JK, Magidson J. Latent class models for classification. Computational Statistics & Data Analysis 2003;41:531–537.

Vulkan N. An economist's perspective on probability matching. Journal of Economic Surveys 2000;14:101–118.

Weiss Y, Simoncelli EP, Adelson EH. Motion illusions as optimal percepts. Nature Neuroscience 2002;5:598–604.

Wichmann FA, Hill NJ. The psychometric function: I. Fitting, sampling, and goodness of fit. Perception & Psychophysics 2001;63:1293–1313. [PubMed: 11800458]

Winkler RL. The consensus of subjective probability distributions. Management Science 1968;15:B61–B75.

Woodbury CB. The learning of stimulus patterns by dogs. Journal of Comparative Psychology 1943;35:29–40.

Young MJ, Landy MS, Maloney LT. A perturbation analysis of depth perception from combinations of texture and motion cues. Vision Research 1993;33:2685–2696. [PubMed: 8296465]

Zhang, H. The optimality of Naïve Bayes; Proceedings of the 17th International FLAIRS Conference; Florida, USA. Menlo Park, California: American Association for Artificial Intelligence; 2004.

Zhu M, Lu AY. The counter-intuitive non-informative prior for the Bernoulli family. Journal of Statistics Education 2004;12:1–10.

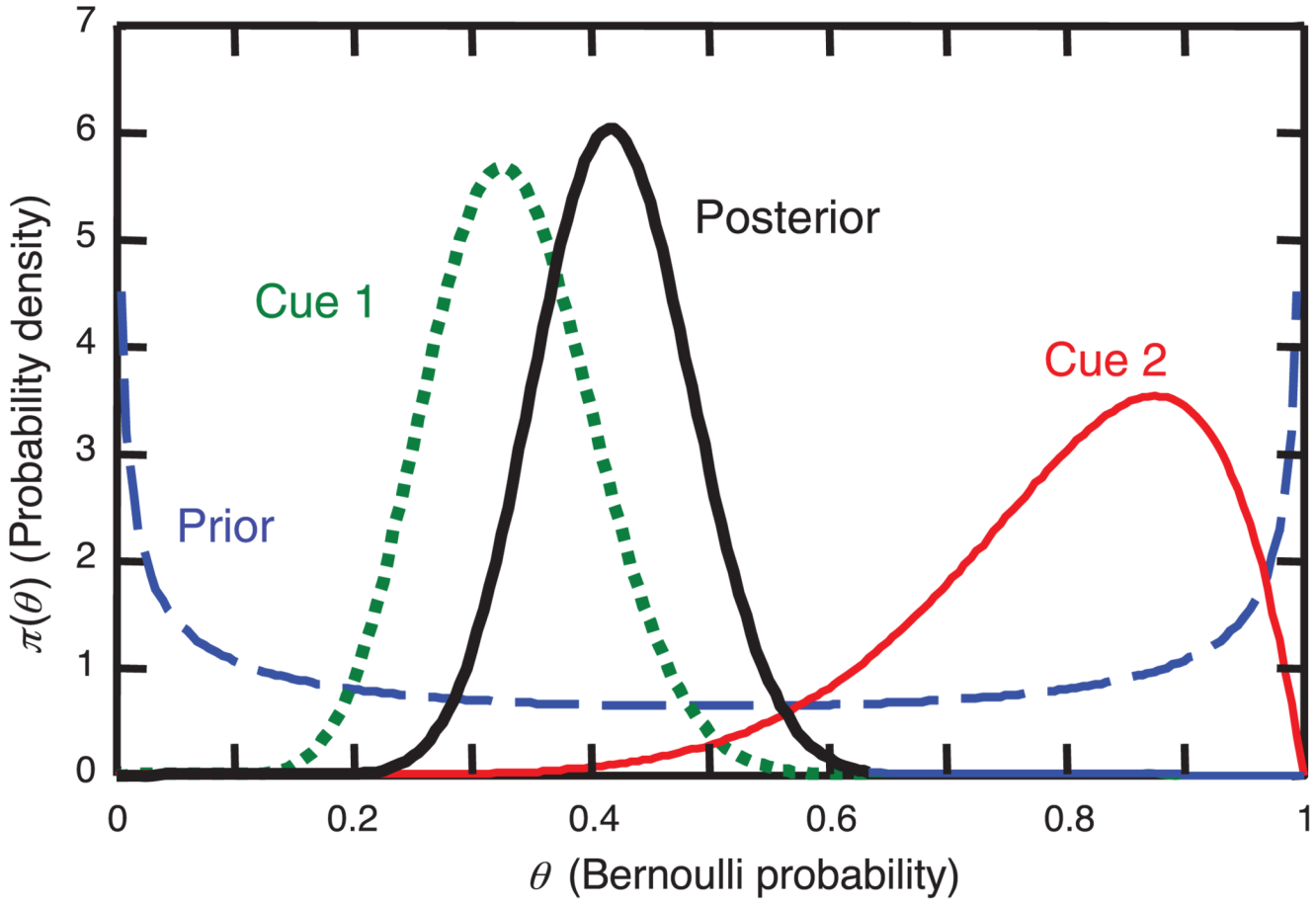

Figure 1.

Example to illustrate the model. Prior to measuring Cue 1 or Cue 2, $\pi(\theta)$ is distributed as beta (0.5, 0.5), shown by the blue dashed line. The cues provide evidence that is used to update the prior, yielding the posterior. The additional belief provided by Cue 1 (dotted green curve) and Cue 2 (solid red curve) can be visualized according to what they would look like if used (separately) to compute a posterior starting with a beta prior that has no reliability, namely $\lim_{\alpha \to 0+} \text{beta}(\alpha, \alpha)$. Using the (θ, n) notation, the prior in the figure is described by $(\theta_0, n_0) = (0.5, 1)$ and the two cues are described by $(\theta_1, n_1) = (0.3\bar{3}, 45)$ and $(\theta_2, n_2) = (0.8, 10)$, respectively. The posterior (solid black curve) is described by $(\hat{\theta}, \hat{n}) = (0.42, 56)$. The noise term has a weight of $n_{-1} = 0$ in this toy example so it does not contribute to the posterior. In this example the prior (Jeffreys' prior, see Zhu & Lu, 2004) contributes to the shape of posterior distribution but has no net effect on the final decision (i.e., whether the expected value of the posterior is greater than 0.5)—as would be the case for any balanced prior.

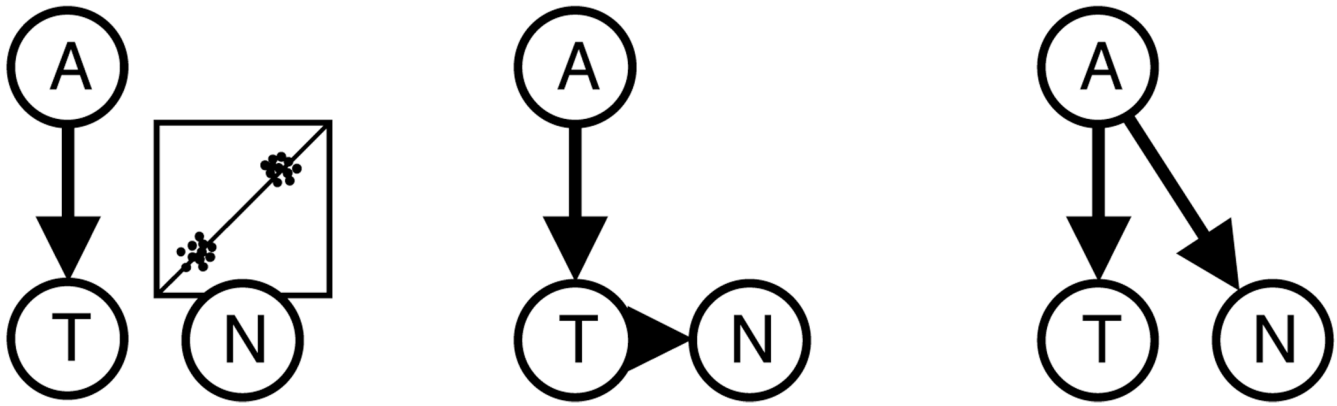

Figure 2.
Causal Bayes net depiction of change in belief during cue recruitment. **A** is a property of the world and **T** and **N** are internal state variables that represent measured cues. *Left*: Only **T** is believed to depend on **A** but a new cue **N** is put into correlation with **T** during training. *Middle*: One way to explain why training causes **N** to be recruited as a cue for **A** is that after training, **N** is believed to depend on **T**. *Right*: Another explanation is that after training **N** is believed to depend on **A**, reflecting an assumption of conditional independence between **T** and **N**. See text for further explanation.

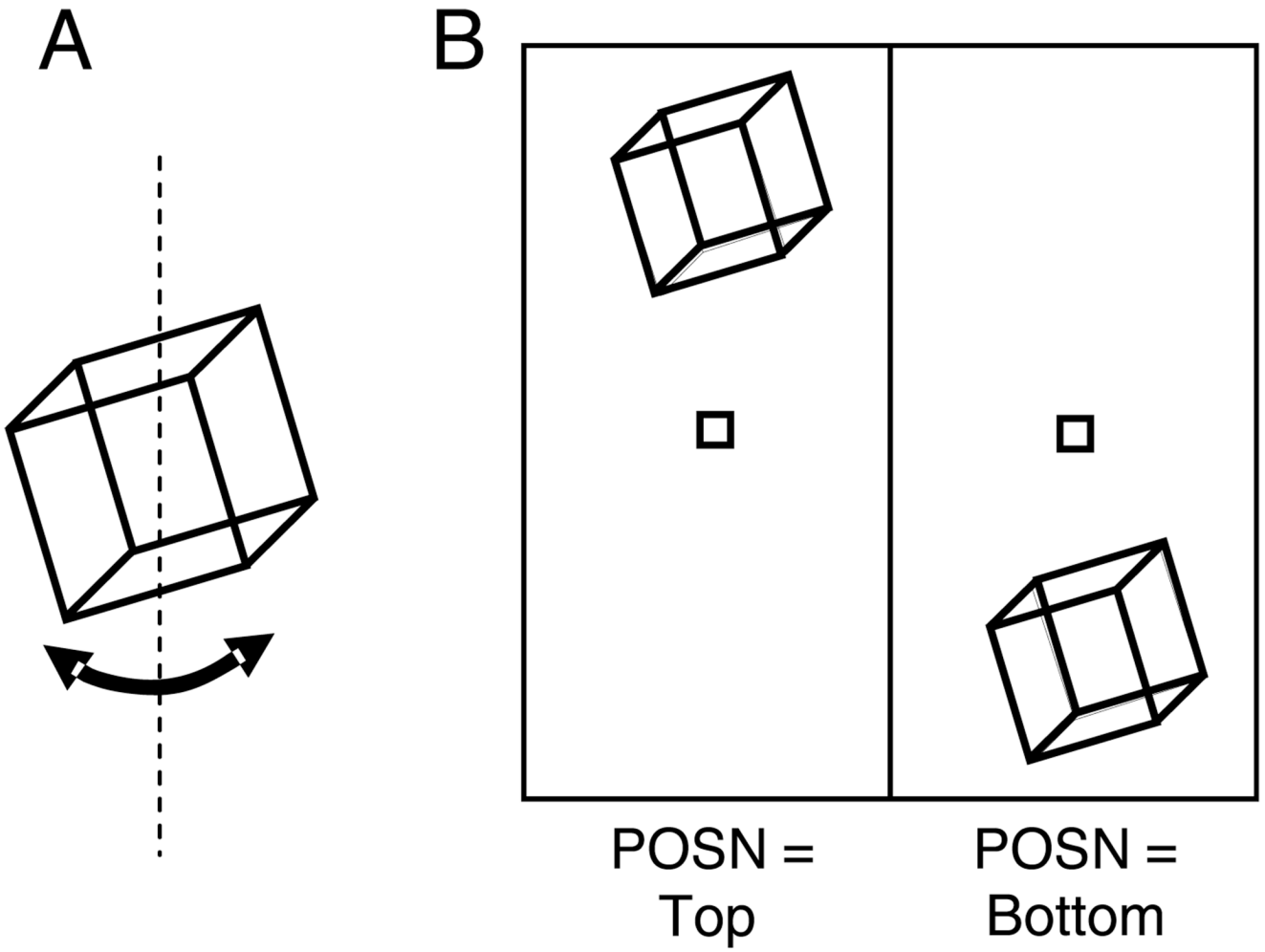

Figure 3.
(A) Depiction of a rotating wire frame (Necker) cube stimulus. It is perceptually bistable and appears to rotate either left or right about a vertical axis. (B) The two values of the POSN cue are shown. The small square at the middle of the display represents a fixation box. A small dot moved either left or right through the fixation box on each trial (determined by random choice) and the trainee pressed a button to indicate whether it moved in the same direction as the front or the back of the cube.

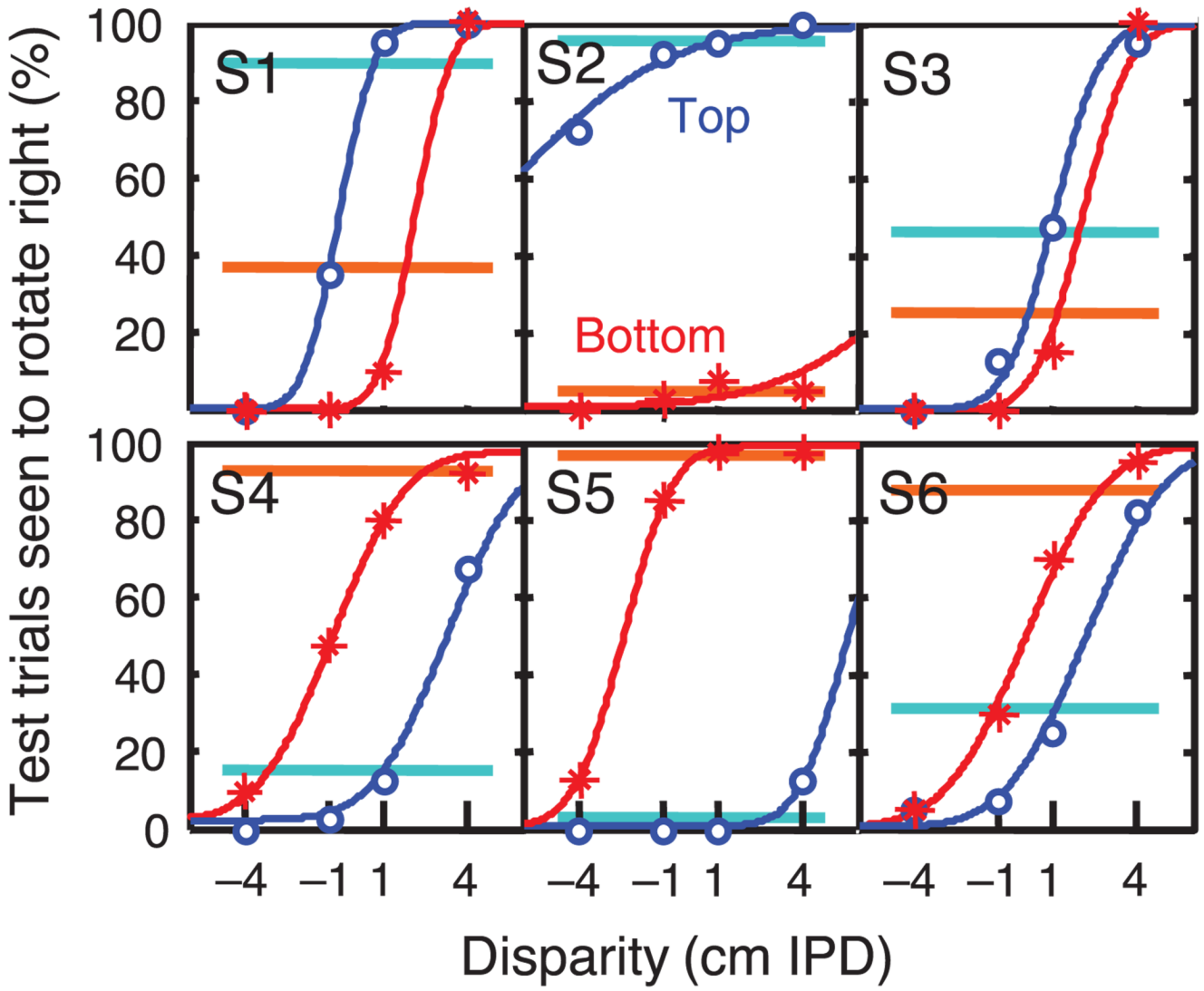

Figure 4.

Test trial data for six trainees. Percent of trials on which the cube was seen to rotate rightward is plotted as a function of POSN on monocular trials (light blue, orange horizontal segments) or as a function of POSN and disparity (blue circles and red stars). The abscissa plots the binocular disparity between rightward- and leftward-moving parts of the cube's image in units of interpupillary distance or IPD; natural disparity was ~6.2 cm. The ordinate plots the percentage of trials (out of 40 monocular trials or 20 binocular trials) in which the cube was seen to rotate rightward. Blue circles and red stars represent data for cubes that were shown above and below fixation, respectively. The blue and red curves are cumulative Gaussians fitted to maximize the likelihood of the data. Trainees who showed a large effect of POSN on binocular trials tended to show a large effect of POSN on monocular trials.

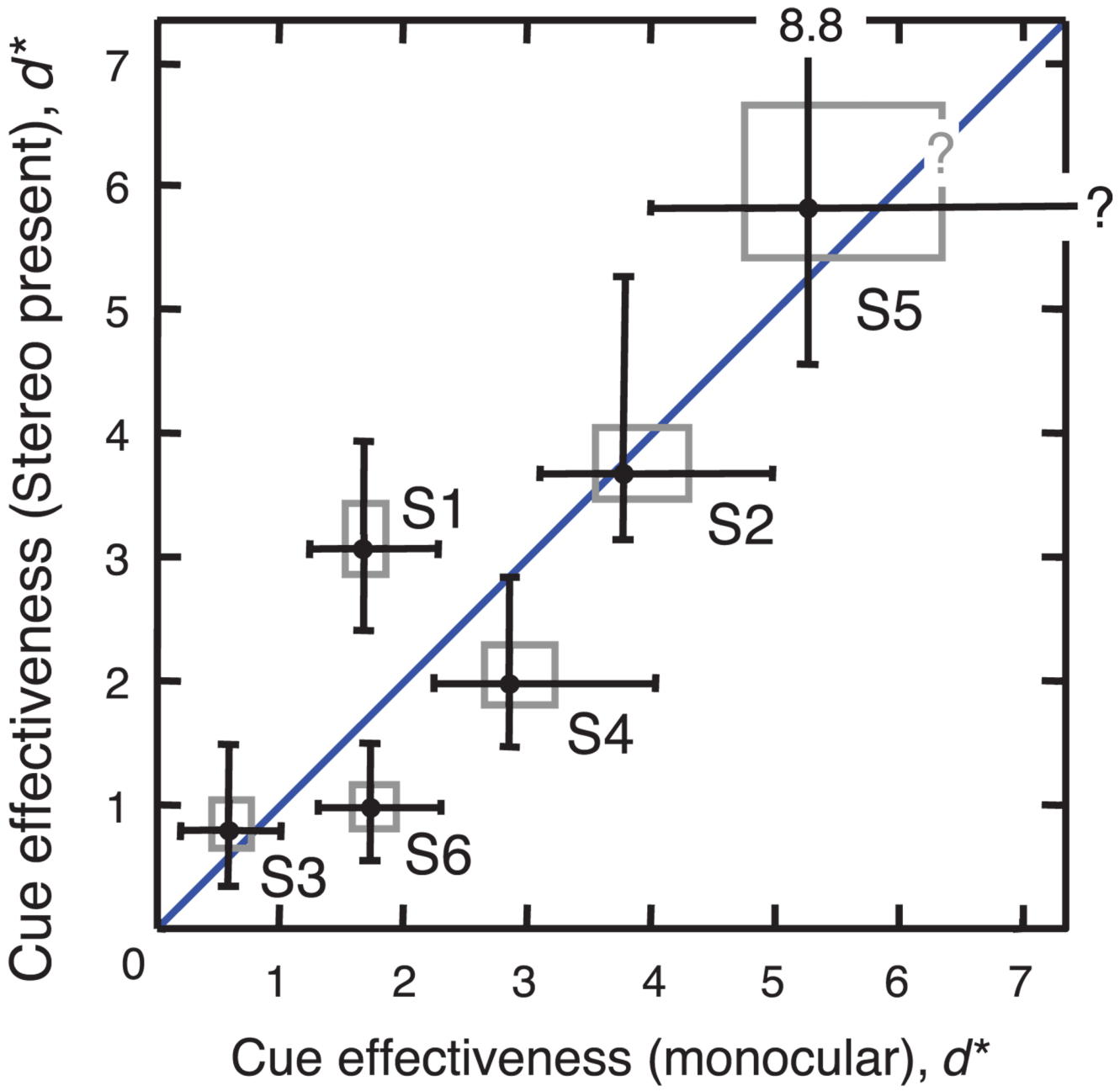

Figure 5.
Comparison of POSN cue effectiveness in monocular and binocular stimuli. Data from the six trainees (S1–S6) in Figure 4 are re-plotted as cue moments (*d**) for the POSN cue. Gray boxes show 50% confidence intervals and black bars show 95% confidence intervals, computed using a bootstrap procedure that resampled data and fitted a probit model as in Figure 4. Confidence intervals for trainee S5 could not be reliably estimated from the data.

Table 1

Symbols and their meanings in MBE.

Sym	Name	Equal to
N	Total number of cues	a whole number
θ_i	ith Bernoulli probability	real no. on [0,1]
n_i	ith cue weight	nonneg. real
ρ_i	Ecological validity of ith cue	$2(\theta_i - 1)$
m_i	Moment or subjective reliability of ith cue (in NEDs)	$n_i\rho_i$
M	Grand moment (in NEDs)	$\sum m_i$
d^*	d^*	change in **M**