How do people build up visual memory representations from sensory evidence? Revisiting two classic models of choice

Maria M. Robinson, Isabella C. DeStefano, Edward Vul, and Timothy F. Brady
University of California San Diego

In many decision tasks, we have a set of alternative choices and are faced with the problem of how to use our latent beliefs and preferences about each alternative to make a single choice. Cognitive and decision models typically presume that beliefs and preferences are distilled to a scalar latent strength for each alternative, but it is also critical to model how people use these latent strengths to choose a single alternative. Most models follow one of two traditions to establish this link. Modern psychophysics and memory researchers make use of signal detection theory, assuming that latent strengths are perturbed by noise, and the highest resulting signal is selected. By contrast, many modern decision theoretic modeling and machine learning approaches use the softmax function (which is based on Luce's choice axiom; Luce, 1959) to give some weight to non-maximal-strength alternatives. Despite the prominence of these two theories of choice, current approaches rarely address the connection between them, and the choice of one or the other appears more motivated by the tradition in the relevant literature than by theoretical or empirical reasons to prefer one theory to the other. The goal of the current work is to revisit this topic by elucidating which of these two models provides a better characterization of latent processes in m-alternative decision tasks, with a particular focus on memory tasks. In a set of visual memory experiments, we show that, within the same experimental design, the softmax parameter β varies across m-alternatives, whereas the parameter d' of the signal-detection model is stable. Together, our findings indicate that replacing softmax with signal-detection link models would yield more generalizable predictions across changes in task structure. More ambitiously, the invariance of signal detection model parameters across different tasks suggests that the parametric assumptions of these models may be more than just a mathematical convenience, but reflect something real about human decision-making.

Introduction

We make choices in virtually every real-world and laboratory task. For example, we decide which cereal we prefer in a supermarket, which color a word is in a Stroop task, or which item is 'old' in a forced-choice memory study. Because decision processes are ubiquitous, there is great value in determining the type of quantitative model that best captures them. To this end, we examine the generalizability of two prominent probabilistic models of choice. The first is a Gaussian signal detection model, which is based on classic Signal Detection Theory (e.g., Wixted, 2020) and Thurstone's law of comparative judgment (Thurstone, 1927). The second is the normalized exponetial model, commonly known as the softmax function (e.g., Bridle, 1990), which is based on Luce's Choice Axiom (LCA) (Luce, 1959) and the ratio of strengths formula (Bradley & Terry, 1952) (for extensive taxonomy of these models see: Townsend & Landon, 1983).

In the current work, we focus on how these two models generalize across different decision-based visual memory tasks in order to better understand the types of computations people use to convert sensory evidence to memory representations to make memory-based decisions. Focusing on the generalizability of these models is key because this allows us to better isolate latent variables of interest (e.g., Navarro, 2021). For illustration, consider a standard forced-choice task in which you are shown an object that you have to remember. Subsequently, when are you tested on your memory, you are shown that object along with one or seven foil objects, where foils refer to objects you were never actually shown. In this simple forced choice task, as more foil items are added you will tend to become less accurate at choosing the object you saw. This follows because your hit rate will decrease as you are presented with more options simply because the probability of you incorrectly choosing a foil will tend to increase when more foils are present (Wickens, 2001). Importantly, if memory conditions are held constant across these decision tasks, the fidelity of your memory for the object you saw should also remain unchanged, regard-

The authors declare no conflict of interest. The study was supported by National Institute of Health Grant 1F32MH127823-01 awarded to MMR and the National Science Foundation Grant BCS-2146988 awarded to TFB.

less of how you are tested on your memory (Swets, 1959). Thus, a key question is what decision model best allows us to assess people's memory strength independently of the decision task we use to test it. In other words, which decision model's parameters are invariant and best generalizes across variations in task structure that affect the decision process but not memory fidelity?

We focus on Signal Detection Theory and Luce's Choice Axiom because they are prominent in different domains, such as decision-making and memory research, but within some domains, there are relatively few comparisons between them. Furthermore, early work that examines the connections between these models (for recent review see: Pleskac, 2015), has yet to be linked to contemporary research questions. We illustrate these points in the context of recent computational modeling research on visual memory.

Our article has the following structure. First, we overview each of the theories and their corresponding models. Second, we outline how early work on the relationship between models based on SDT and LCA applies to contemporary research on visual memory and describe a critical test that we used to discriminate between them. Finally, we discuss our findings and their relevance for theorizing about decision processes within and outside of visual memory tasks.

Signal detection theory

The application of SDT to the study of sensory and cognitive process comes from the tradition of perceptual psychophysics (e.g., Green, Swets, et al., 1966), which highlights the relationship between sensory signals that must be used to make a decision, and the physical and neural noise that perturbs them before a decision is made (Wickens, 2001). Over the years, SDT has been used in other domains, such as memory research (e.g., Wixted, 2007), to provide a detailed description of decision processes in detection and discrimination tasks by postulating latent memory-strength signals that are perturbed by noise. The two core assumptions of signal detection models is that when faced with making a decision, the conceivably rich and multi-dimensional representation of each alternative is collapsed down into a scalar value –the decision variable– and that the decision variable invoked by a particular alternative is probabilistic. Jointly, these assumptions capture the mainstream view that there are internal and external sources of noise that corrupt sensory and memory signals (e.g., Dosher & Lu, 1998). For instance, in the memory domain, a familiar object, such as a backpack, will produce a decision variable of some magnitude with respect to some task, such as a familiarity signal for a recognition task. The decision variable produced by observing a backpack will vary from one instance to another due to variation in external circumstances, such as its lighting and vantage point, as well as fluctuations of internal states, such as memory, attention and motivation.

Because decision variables in this view are seen as random variables, it is common to postulate a specific probability distribution over them (although see: Kellen et al., 2021). While in some low-level perceptual domains, great care has been taken to characterize the functional form of this distribution, and thus the form of the psychometric function (e.g., Green, Swets, et al., 1966), in most applications such fidelity is unattainable and researchers simply assume that decision variables are normally distributed. Thus, historically, the normality assumption common in SDT is made primarily for convenience (Wickens, 2001). Furthermore, in contemporary modeling work it is often treated as an auxiliary assumption that does not have a theoretical justification (Kellen et al., 2021; Rouder et al., 2010). To preview our analysis and results, we show that the Gaussian parameterization of signal detection models is not merely ancillary. Instead, its use can have a principled theoretical basis that formalizes how sensory signals are converted to decision variables. We discuss this point in depth when reviewing the mathematical link between the Gaussian signal detection and softmax model.

Finally, most mainstream signal detection models postulate that, while decision variables are probabilistic, the decision making process is deterministic (for exceptions see, e.g., Benjamin et al., 2009). That is, once decision variables are sampled from their probability distributions, choices are made deterministically by comparing the decision variables to one another, or to a fixed decision criterion. Next we describe how these principles are used to explain performance in mainstream detection and discrimination tasks.

SDT for detection and discrimination tasks

In detection tasks the observer responds by indicating the presence or absence of a target stimulus. The classic Gaussian signal detection model posits that this decision is made by collapsing the rich stimulus representation down into a single decision variable and then comparing this decision variable X against a fixed decision threshold C. Accordingly, the probability of responding that a target is "Present" on target present and absent trials is given by Equations 1 and 2, respectively:

$$P(\text{'}Present\text{'} \mid Present) = P(X_T > C), \tag{1}$$

$$P(\text{'}Present\text{'} \mid Absent) = P(X_F > C). \tag{2}$$

In Equation 1 X_T denotes the decision variable elicited by the target stimulus, which is a random variable sampled from a normal distribution with free parameters, mean $\mu > 0$ and variance σ^2: $X_T \sim \mathcal{N}(\mu, \sigma^2)$. A common assumption is that, on average, decision variables on target present trials will be of greater magnitude than on target absent trials, and it follows that their mean will also be greater. Therefore, with no loss in generality, the mean and variability of the decision

variable elicited by foil items, X_F in Equation 2, on target absent trials is set to 0 and 1, respectively: $X_F \sim \mathcal{N}(0, 1)$.

Unlike in detection tasks, in forced-choice discrimination tasks the target is always shown and an observer must select it out of a set of n alternatives. Classic signal detection models postulate that this selection process involves computing the maximum of a set of n independent random variables corresponding to the decision variables invoked by each of the stimuli: X_i. More precisely, the probability of identifying a given item i as the target is the probability that the magnitude of the decision variable generated by the target X_i exceeds the decision variables generated by each of the $n-1$ foil items X_j for $j \neq i$:

$$P(ID(i)) = P(\forall j \neq i : X_i > X_j). \tag{3}$$

This general expression can be written out for the special cases of correct choices, when X_i corresponds to the target ($i = 1$), and incorrect choices, when $i \neq 1$. For correct choices, or Target Identifications, X_i is the target ($X_i = X_1 = X_T$) and all X_js are foils, thus $X_i \sim \mathcal{N}(\mu, \sigma^2)$, and $X_j \sim \mathcal{N}(0, 1)$. For incorrect choices, or Foil Identifications, the target is one of the X_js while X_i and the remaining X_js are foils. For both of these special cases, we can rewrite the general expression:

$$X_1 = X_T \sim \mathcal{N}(\mu, \sigma^2) \tag{4}$$

$$X_{2\ldots n} \underset{\text{iid}}{\sim} \mathcal{N}(0, 1) \tag{5}$$

$$P(ID(\text{Target})) = P(X_1 > \max(X_{2\ldots n})), \tag{6}$$

$$P(ID(\text{Foil})) = \sum_{i=2}^{n} P(X_i > \max(X_{1\ldots n \setminus i})). \tag{7}$$

Luce's choice axiom

Luce's Choice Axiom (LCA) comes from the decision theory tradition, rather than psychophysics, and is predated by the ratio of strengths formula for pairwise choices (Bradley & Terry, 1952) (for empirical tests and extended discussion of these models see: Townsend & Ashby, 1982; Townsend & Landon, 1983). Unlike SDT, the LCA framework is silent about the mechanisms of detection and discrimination processes. Instead, it consists of a set of axioms that impose "plausible constraints" on choice probabilities.

The central axiom is called Independence from Irrelevant Alternatives and states that the probability of choosing one alternative over another should not change if irrelevant alternatives are added or taken away. Under this view, response probabilities for each alternative are computed by dividing each response strength by the sum of all response strengths in the set. For instance if a is one alternative out of a larger set T, the probability of choosing a out of S is

$$P(a, S) = \frac{\phi(a)}{\sum_{z \in S} \phi(z)}, \tag{8}$$

where ϕ is a response strength function. Note that independence from irrelevant alternatives follows directly from this formula because the odds of choosing a over a different alternative $b \in S$ remains the same, even if we consider a larger set of alternatives T where $S \subseteq T$. That is, for $0 < P(x) < 1$,

$$\frac{P(a, S)}{P(b, S)} = \frac{P(a, T)}{P(b, T)} = \frac{\phi(a)}{\phi(b)}. \tag{9}$$

Equation 8 also implies that the function ϕ lies on a ratio scale. That is, assume there exists another function ϕ' that satisfies the equality

$$\frac{\phi(a)}{\phi(b)} = \frac{\phi'(a)}{\phi'(b)}. \tag{10}$$

Substituting 1 for $\phi(b)$ and $\tau > 0$ for $\phi'(b)$ yields $\tau\phi(a) = \phi'(a)$, showing that the scale ϕ is unique up to multiplication by a positive constant (proof adapted from: Krantz et al., 1971). This entails that the response function ϕ lies on a ratio scale, an important and rare property of psychological metrics (Falmagne & Doble, 2015).

Finally, note that in order for choice probabilities in Equation 8 to be restricted between zero and one, response strengths should be constrained to be non-negative. One way to impose this constraint is to parameterize the Luce choice model with an exponential function, such that

$$P(a, S) = \frac{e^{\phi(a)}}{\sum_{z \in S} e^{\phi(z)}}. \tag{11}$$

This formulation of LCA is equivalent to the exponential form of the multinomial distribution and the softmax function (Bridle, 1990), which is routinely used in econometrics (McFadden, 1980), machine learning (Murphy, 2012) and reinforcement learning (Sutton & Barto, 2018).

LCA for detection and discrimination tasks

Through the lens of LCA, performance in detection and discrimination tasks is not determined by random decision variables but by fixed response strengths. In detection tasks, assume that β denotes response strength generated by the target stimulus[1] and V denotes a bias parameter for reporting the stimulus is absent. Then, on target present trials, the probability of correctly responding target present is

$$P(\text{'}Present\text{'} \mid Present) = \frac{e^{\beta}}{e^{\beta} + e^{V}}. \tag{12}$$

[1] Technically, β denotes how response strengths are weighted. More precisely, as β increases responses become more deterministic, such that alternatives with higher response strengths receive more weighting and are more likely to be chosen. However, since we assume that foil items yield zero response strength in this exposition, we equate β with response strength of the target stimulus.

On target absent trials, the probability of incorrectly responding target present is determined by the response strength generated by the foil, which is zero. Thus, the probability of incorrectly responding target present on target absent trials is

$$P(\text{'Present'} \mid Absent) = \frac{1}{1 + e^{V}}. \tag{13}$$

Note that the formulas for choice probabilities in Equations 12 and 13 are formally equivalent to a logistic cumulative distribution (Suppes & Krantz, 2007), a special case of the softmax function for binary choices.

Extending this logic to discrimination tasks with n alternatives uses the standard assumption that the response strength generated by the target and $n - 1$ foils is equal to β and zero, respectively. Accordingly, the probability of correctly selecting the target is

$$P(ID(\text{Target})) = \frac{e^{\beta}}{e^{\beta} + n - 1}, \tag{14}$$

and the probability of incorrectly selecting a foil item is

$$P(ID(\text{Foil})) = \frac{n - 1}{e^{\beta} + n - 1}. \tag{15}$$

Connections between SDT and LCA

Due to their distinct origins and distinct mathematical instantiations, models based on SDT and LCA may seem extremely different from one another. However, the Gaussian signal detection and softmax models turn out to be close approximations in some tasks. More precisely, in detection tasks, the connection between these models follow simply from the fact that the logistic distribution approximates the normal distribution and vice versa (Treisman & Faulkner, 1985). This entails that the LCA for binary choices is equivalent to a signal detection model with a logistic parameterization, which closely approximates the Gaussian signal detection model. Thus, in detection tasks LCA and Gaussian signal detection models are closely related.

In discrimination tasks with more than two alternatives the Gaussian signal detection and softmax model no longer approximate each other. The relationship between these two models breaks down in m-afc tasks (where $m > 2$) because the distribution of maximums of normally distributed variables is not a normal distribution. However, it is possible to establish an equivalence between the two models by dropping the normality assumption in the signal detection model. Holman and Marley, 1974 as well as Yellott Jr, 1977 showed that, if decision variables in the signal detection model have a Type 1 extreme value Gumbel distribution for the maximum (Gumbel, 1954), than the signal detection model is mathematically equivalent to the Luce model for any number of alternatives (m) in an m-afc task.We provide our own proof of this result in the Appendix.

In the current context, the major implication of this result is that comparing the softmax model to the Gaussian signal detection model can be recast as a comparison of two different parameterizations of the signal detection model, that is, a signal detection model with a Gumbel versus a Gaussian parameterization. As we discuss next, these two parameterizations have an important conceptual basis because they describe different ways of translating sensory evidence into decision variables.

Processing implications of a Gaussian versus Gumbel signal detection (softmax) model

A common assumption is that the Gaussian parameterization of signal detection models is made for mathematical convenience and does not have a theoretical basis (e.g., Kellen et al., 2021). However, early work by Thompson and Singh, 1967 provides one principled justification for using a normal distribution to model decision variables. These researchers noted that each time we observe a stimulus, it produces a sensory response of some variable magnitude. For instance, through the lens of contemporary population coding neural models, these sensory responses can be conceived of as distributed patterns of activation in populations of neurons (e.g., Averbeck et al., 2006).

If this large number of sensory signals (e.g., patterns of activation across a population) are pooled together by summing or averaging to compute decision variables, then in accordance with the Central Limit Theorem, decision variables will be normally distributed. In contrast to the Gaussian, the Gumbel distribution is an extreme value distribution used to model the maximum of a set of random variables (Gumbel, 1954). Thus, a signal detection model with a Gumbel parameterization is most consistent with the view that, rather than pooling, the observer takes the maximum of sensory signals to compute decision variables. Figure 1 depicts these predictions by showing how a stimulus produces a neural response profile that consists of a set of tuning functions (colored distributions), and how these neural responses can be converted to a single decision variable through the lens of each model.

Together, a test between these models can be recast as a test of two different signal detection models. To further motivate the comparison of these two models, we underscore that there is extensive support for signal detection theory as a general theory in the memory domain (for recent overview see: Wixted, 2020) using diverse methods, including Receiver Operating Characteristics analysis (e.g., Robinson et al., 2020; Williams et al., 2022; Wixted, 2007), and a novel critical test which rests on minimal assumptions (Winiger et al., 2021) While some authors reported evidence for alternative models under some conditions (e.g., Balakrishnan, 1999; Rouder et al., 2008), follow-up work suggests that these results were spuriously driven by either restricted model assumptions, or non-diagnostic data and inadequate

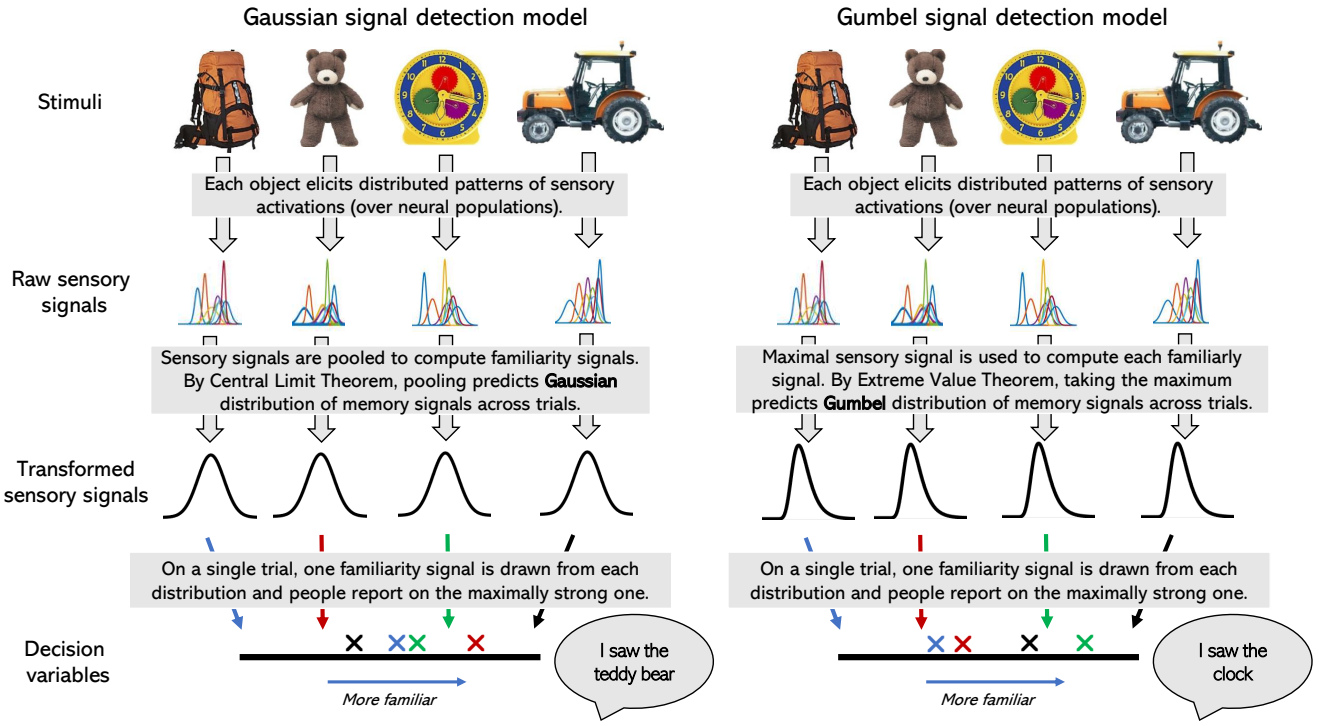

Figure 1

Processing implications of the Gaussian versus Gumbel parameterization of the signal detection model. By Central Limit Theorem, the Gaussian signal detection model entails that observers convert sensory evidence (depicted with colored distributions) evoked by a stimulus (such as backpack) to decision variables via pooling. The Gumbel signal detection model, which is formally equivalent to the softmax model, entails that observers convert sensory evidence to decision variables by taking the maximum of the sensory signals.

metrics of model fit (Mueller & Weidemann, 2008; Robinson et al., 2022). Moreover, recent modeling work in the visual memory domain indicates that a signal detection model constrained by psychophysical scaling methods outperforms all extant alternative models both in fit and generalization (Schurgin et al., 2020). Thus, classic and contemporary modeling work demonstrates robust evidence for signal detection models of memory. Our work builds on this literature by highlighting that the parametric assumptions of signal detection models are not merely ancillary, but can have different implications for how we think observers convert rich sensory or memory evidence to decision variables when making memory-based decisions.

Critical test: Parameter invariance across changes of m in m-afc tasks

We compared the Gaussian signal detection and softmax model by examining which model's parameters (d' in SDT; β in LCA/softmax) are invariant across variations in the number of alternatives presented at test in an m-afc task. Our test rests on the assumption that, everything else being equal, the way in which observers compute decision variables should be invariant across changes in m-afc. This assumption aligns with the broader view that model parameters that generalize across task structures may also provide better approximations of latent cognitive processes (Busemeyer & Wang, 2000).

We note that a similar test was used in an auditory memory task in an early study by Treisman and Faulkner, 1985. These authors reported evidence for the Gaussian signal detection model, however, their results were somewhat ambiguous. Mainly, they found that variations in m-afc produces decreases in d' and increases β, parameters in the Gaussian signal detection and softmax model, respectively. The researchers interpreted this as evidence for the Gaussian signal detection model because they reasoned that increasing the number of alternatives in the auditory task may increase memory load and hurt performance, but not improve it. However, while the finding that d' decreases with m may be more psychologically plausible, it does not demonstrate that parameters of this model are invariant with m because m is confounded with memory load. Furthermore, this study only

used data from 6 participants and may have been underpowered. As we discuss next, one of our goals was to address both of these methodological limitations and perform a more general test of the two theories.

Application to visual memory

We ran a new set of experiments that extends the critical test of Treisman and Faulkner, 1985 to the visual memory domain. The first reason we used a visual memory task is because this allows us to present all *m*-afc alternatives visually, instead of having participants maintain these in working memory. Accordingly, this study design minimizes differences in memory load across *m*-afc task, addressing the core limitation of the Treisman and Faulkner experiment and providing a strong test bed of parameter invariance. We also increase the number of participants in our experiments to ensure that our studies are sufficiently powered.

Another motivation for extending this test to the visual memory domain is because a comparison between these models has direct relevance for contemporary models of visual memory. That is, both the Gaussian signal detection and softmax models have been used in recent modeling work as response functions that capture how people make decisions in *m*-afc visual memory tasks (Oberauer & Lin, 2017; Schurgin et al., 2020). However, these models have not been empirically compared with critical tests, and the processing implications for understanding how people compute decision variables in visual memory tasks have not been discussed. Finally, there is much recent interest in instantiating human visual memory models using neural network architectures (e.g., Bates et al., 2023; Brady & Störmer, 2020; Hedayati et al., 2022) that routinely use the softmax as a response function (Murphy, 2012); it remains unclear whether this provides the best approximation of how humans map latent states to memory judgments. Our goal is to fill these gaps by comparing these models in a set of visual memory experiments. To this end, we ran two experiments in which we varied the structure of the stimulus space, the dimensionality of stimuli and the presentation format to ensure that our results were robust across different processing domains and theoretical assumptions.

Experiment 1: Memory for simple features

Experiment 1 was designed to test the signal detection and softmax models in a multiple alternative forced choice visual working memory task with simple features (color). The central comparison involves examining which model's parameters are invariant across changes in the number of alternatives in *m*-afc tasks.

Methods

Participants Participants (n = 31) were undergraduate student volunteers, at the University of California, San Diego, who participated in the study for course credit. All participants were at least 18 years old, reported normal or corrected-to-normal vision, and provided informed consent. All experiments were approved by the Institutional Review Board at the University of California, San Diego.

Our predetermined sample size was n = 30. This sample size is a conservative bound for detecting a medium effect size (d_z = 0.6) with 90% power and α = .05 significance criterion. We collected participant data until our sample size reached n = 30 based on our exclusion criteria. Consistent with our standard lab practice, we excluded trials with reaction times less than 100 ms or greater than 5000 ms (average proportion of 3% per participant). We excluded participants who had more than 10% of trials excluded, or who whose performance was at chance in any of the four conditions (one participant).

Stimuli Stimuli were colored circles. Colors were drawn from the CIE L*a*b* color space, centered in the color space at (L = 54, a = 21.5, b = 11.5) with a radius of 49 (from Schurgin et al., 2020).

Procedure On each trial, participants were shown four circles and instructed to remember their colors and spatial locations. The minimum distance (along the color circle) between each circle in the memory array was 30 degrees. The memory array was shown for 1,000 ms. After a brief retention interval (800 ms), participants were shown a spatial cue that probed one of the four circles shown in the memory array.

Participants were instructed to use a discretized color wheel to report on their memory for the probed circle. The discretized color wheel consisted of 2, 4, 8, or 16 colors, which were spaced either 180°, 90°, 45°or 22.5°apart in color space, respectively. Participants were instructed to click on the color that they thought best matched the color of the probed circle. One of the colors always matched the color of the probed circle, whereas the others did not. There were a total of 500 trials in the experiment (125 trials per *m*-afc condition), and each experimental session lasted approximately 50 minutes.

Analysis

Participants' responses were converted to errors by taking their distance along the color wheel from the correct answer, where the correct response is centered at zero. Rather than assuming that all foils are processed independently and elicit zero signal regardless of their similarity to the shown color, we fit models to the full distribution of errors under the assumption that the latent memory strength of each alternative scales with the psychophysical similarity to the remembered

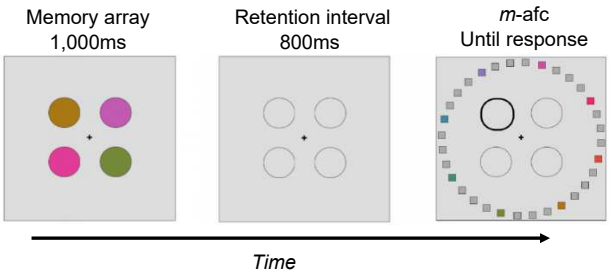

Figure 2

Example trial from Experiment 1. On each trial participants were shown a memory array with four colored circles. The memory array was presented for 1,000 ms and followed by an 800 ms retention interval. After a retention interval, participants were shown a self-report screen with 2, 4, 8, or 16 equally spaced colors, and the other positions filled with gray "filler" squares. One of the colors presented at test was always shown on that trial, and the remaining colors were not. Participants had to click one the colors to indicate which color was at the cued position on this trial. Responses were not speeded. The pictured trial shows an 8-AFC test with 8 colors presented at test, and the correct answer is the yellow color on the bottom of the response wheel, as this matches the color presented in the top left location, which is the cued location on this trial.

item. This assumption aligns with classic feature matching models of memory (e.g., Clark & Gronlund, 1996), as well as more recent work on visual memory (Schurgin et al., 2020). More precisely, both theoretical frameworks predict that foils that are more similar to the target have stronger latent memory strengths than those that are less similar. This fitting procedure used psychophysical similarity values obtained by Schurgin et al., 2020. These deviation values were fit with a Gaussian signal detection model and a softmax model to estimate parameters d' and β, respectively. Models were fit separately to each participant's data and fitting was implemented in MATLAB using maximum likelihood estimation (MLE).

Our goal was to determine which model's parameters are invariant across the *m*-afc manipulations. We tested this by comparing the relative fits of the Gaussian signal detection and softmax model when parameters d' and β, respectively, were fixed across all *m*-afc conditions[2]. This analysis provides insight into which model best accommodates the data if we assume that its parameters are invariant across manipulations of *m*-afc. Since both the signal detection and softmax models have the same number of parameters, we used the log likelihood (LL) to compare models (note that larger values of the LL indicate superior fit). These values were compared at the level of individual participants using a paired *t*-test.

We also implemented three secondary analyses that complement our critical test. In the first complementary analysis, we ran a manipulation check that memory fidelity, which is captured via model parameters, did not decrease as a function of alternatives in *m*-afc. In the second complementary analysis, we evaluated the relative flexibility of both models because prior work suggests that variants of the ratio of strengths formula can be extremely complex relative to alternative models with the same number of parameters (Myung & Pitt, 1997; Pitt et al., 2002; Townsend & Ashby, 1982). To compare these models on their flexibility, we assessed the relative fits when parameters varied freely across *m*-afc conditions. This analysis provides insight into whether (as expected) the invariance of parameters of these models, as opposed to their functional form, yields better fits to data in our primary analysis when parameters are fixed across *m*-afc conditions. That is, we expect that if these models are matched on their flexibility, they should yield comparable fits to the data when parameters vary freely across experimental conditions. In the third complementary analysis, we compared the standard deviation of parameters across *m*-afc conditions when we allowed these to vary freely across *m*-afc conditions. We expect variability of parameters across *m*-afc conditions to be smaller in the model that provides the best fit to data when parameters are fixed across *m*-afc conditions.

Results

Manipulation check

Panel A of Figure 3 shows parameter estimates from each model when they were allowed to vary freely across *m*-afc experimental conditions. Note that one critical assumption of our analysis is that by making $m-afc$ alternatives visible to participants we do not increase memory load with $m-afc$. To check for this, we examined whether parameter estimates decreased systematically as a function of the $m-afc$ manipulation. We found that they did not. Specifically, in the Gaussian signal detection model, average d' equaled 1.79 (SEM = .17), 1.86 (SEM = .16), 1.98 (SEM = .14) and 1.84 (SEM = .16) in the 2, 4, 8, and 16 *m*-afc conditions, respectively. In the Gumbel signal detection model, average β estimates equaled 2.42 (SEM = .34), 2.5 (SEM = .23), 2.85 (SEM = .19) and 3.07 (SEM = .24) in the 2, 4, 8, and 16 *m*-afc conditions, respectively. Together, through the lens of both models we did not find that performance decreased systematically as we increased the number of alternatives at

[2]Model comparisons are essential because in *m*-afc tasks models cannot be compared by simply examining the distributions of errors. For instance, the maximum rule Gaussian signal detection model does not predict perfectly Gaussian distribution of errors because distribution of the maximums of $m > 2$ variables is slightly skewed.

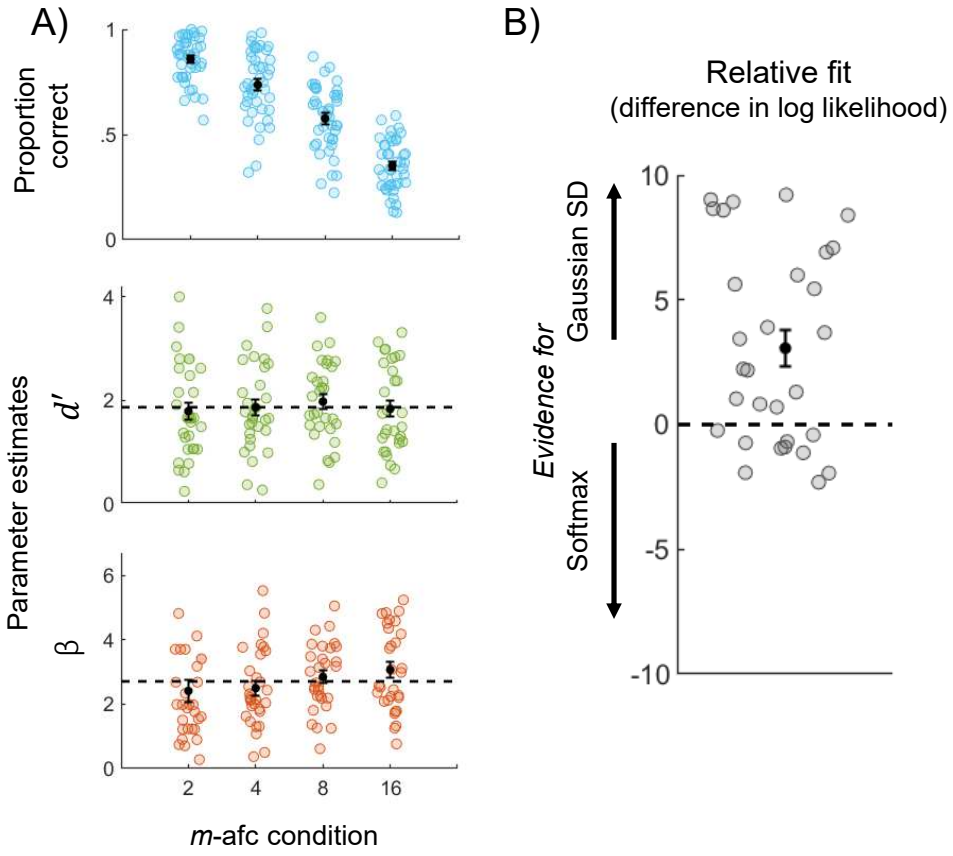

Figure 3

Model fitting and comparison results from Experiment 1. Panel A shows the difference in log likelihood between the Gaussian signal detection and softmax models when we fixed each model's parameters across m-afc conditions. Positive values indicate support for the Gaussian signal detection model, negative values indicate support for the softmax (or Gumbel signal detection model), and values at zero indicate equal support for both models (denoted with the black dotted line). Panel B shows participants raw proportion correct at each m-afc condition (top) as well as the parameter estimates obtained from fitting the Gaussian SDT model (middle) and softmax model (bottom) separately to the full error distributions from each m-afc condition. In each figure, the black dot and error bar denote the average and standard error of the mean across participants within each condition. The black dotted line in Panel B, denotes the mean across participants and condition. The fact that estimates of d′ are more stable than estimates of β across m-afc conditions is consistent with the model comparison favoring the Gaussian SD model.

test.

Model flexibility

Importantly, based on LL, we found that the Gaussian signal detection ($\bar{X} = -560$; $SEM = 31$) and softmax ($\bar{X} = -560.01$; $SEM = 31$) models fit the data comparably when parameters were free to vary across *m*-afc conditions ($t(29) = .19$, $p > .84$; $d_z = 0.04$). This analysis provides convergent support for the conclusion that the superior fit of the Gaussian signal detection model when its parameters are fixed across *m*-afc conditions, reflects its superior capacity to capture invariants across *m*-afc conditions, as opposed to this model having a more flexible functional form.

Critical test for parameter invariance

Panel B of Figure 3 shows the difference in log likelihood (LL) between the Gaussian signal detection and softmax model when parameters d' and β, respectively are fixed across *m*-afc conditions. Positive and negative values indicate support for the Gaussian signal detection and softmax model, respectively, whereas values near zero indicate equal support for both models. We found that the LL was significantly higher for the Gaussian signal detection ($\bar{X} = -562.2$; $SEM = 31$) than the softmax ($\bar{X} = -565.3$; $SEM = 31$)

model ($t(29) = 4.26$, $p < .001$; $d_z = 0.77$). Average parameter estimates in the fixed models were $d' = 1.86$ ($SEM = .15$) and $\beta = 2.69$ ($SEM = .22$).

Based on the standard deviation of parameters across m-afc conditions, we also found that there was significantly less variability in parameters in the Gaussian signal detection ($\bar{X} = .21$; $SEM = .02$) than the softmax model $\bar{X} = .55$; $SEM = .08$) when these were allowed to vary freely ($t(29) = 5.26$, $p < .001$; $d_z = 0.96$). Thus, the d' parameter of the Gaussian signal detection model was more stable across m-afc conditions than the β parameter of the softmax model.

Together, our results provide support for the Gaussian signal detection over the softmax model. That is, we find that the Gaussian signal detection model does a better a job at capturing invariance of decision latent processes across m-afc conditions, and that these effects are not due to differences in model flexibility.

Experiments 2: Memory for real-world objects

The goal of Experiment 2 was to examine the generalizability of our modeling results. To this end, in Experiment 2 we modified both the stimuli and presentation format. More precisely, we had participants remember real-world objects instead of simple features and presented stimuli sequentially instead of simultaneously. Another advantage of using real-world objects instead of colors as stimuli, is that the real-world object stimulus space in unconstrained. This entails that we can select a larger number of foils that are completely dissimilar from the target and, therefore, compare models without relying on additional assumptions about how participants process psychophysically similar foils.

Methods

Participants Participants ($n = 31$) were undergraduate student volunteers, at the University of California, San Diego, who participated in the study for course credit. As in Experiment 1, we collected participants until we reached a final sample size of ($n = 30$). Exclusion criteria were the same as those used in Experiment 1. We excluded an average of 4% of trials per participant, and one participant.

Stimuli Stimuli were photos of real-world objects taken from Brady et al., 2008. All objects were from different categories.

Procedure On each trial, participants were shown a sequence of five unique photos of real-world objects. Each object was presented for 300 ms, and the interstimulus interval was 100 ms. The sequence of objects was followed by a retention interval that lasted 800 ms.

At memory test, participants were shown 2, 4, or 8 objects and were instructed to click on the object that was shown in the sequence on that trial. We include 3 instead of 4 m-afc conditions because trials with sequential presentation are longer and we wanted to ensure that the experimental session did not run over the 50 minute time limit, while maintaining a sufficiently large number of trials per condition. One of the objects always matched an object shown on that trial sequence, whereas the others were completely novel objects that were only shown once throughout the entire experimental session. There were a total of 210 trials (70 trials per m-afc condition), and each experimental session lasted approximately 50 minutes.

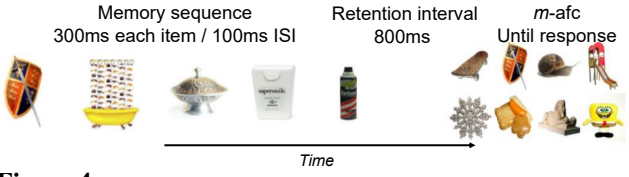

Figure 4

Example trial sequence from Experiment 2. On each trial participants were shown a sequence of five unique photos of real-world objects. Each object was presented for 300ms, with an inter-stimulus interval (ISI) of 100ms. The object sequence was followed by an 800ms retention interval, and then a self-report screen. The self-report screen showed 2, 4, 8 objects. One of the objects was always an object shown on the trial sequence, and the remaining objects were foils from different categories, that were not shown again during the experimental session. Participants had to click which object they had seen on that trial. Responses were not speeded.

Analysis

In this experiment all stimuli, including targets and non-targets in the memory array, as well as foils were chosen randomly with the constraint that they came from different categories (as in Brady et al., 2008). Accordingly, there is no structure to the error distribution as a function of similarity, and analyzing the accuracy data and error distributions yields identical results. Thus, the analysis was identical to the one used in Experiment 1, with the exception that we fit models to proportion correct alone rather than the complete error distribution with no loss in generality.

Results

Manipulation check

Panel A in Figure 5 shows parameter estimates from each model when these varied freely across experimental conditions. In the Gaussian signal detection model, average d' equaled 1.42 ($SEM = .12$), 1.49 ($SEM = .13$) and 1.61 ($SEM = .13$) in the 2, 4 and 8 m-afc conditions, respectively. Average β estimates equaled 1.71 ($SEM = .16$), 1.96 ($SEM = .17$) and 2.30 ($SEM = .18$) in the 2, 4 and 8 m-afc conditions, respectively. Again, through the lens of

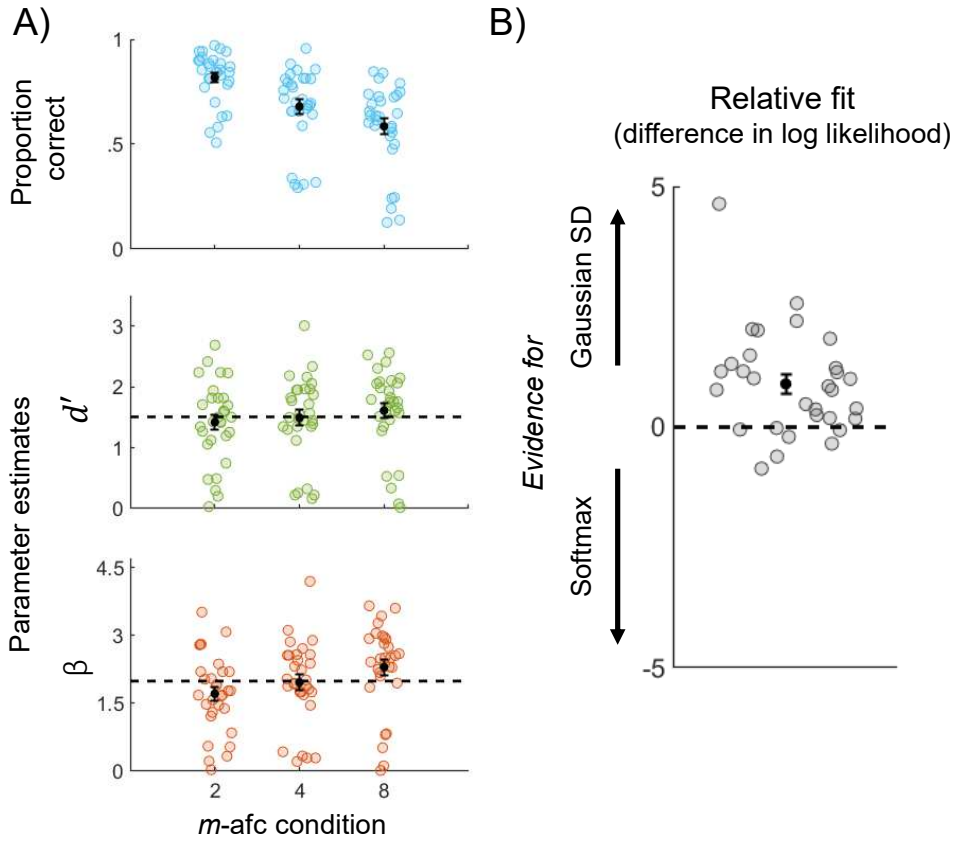

Figure 5

Model fitting and comparison results from Experiment 2. Panel A shows the difference in log likelihood between the Gaussian signal detection and softmax models when we fixed each model's parameters across m-afc conditions. Positive values indicate support for the Gaussian signal detection model, negative values indicate support for the softmax (or Gumbel signal detection model), and values at zero indicate equal support for both models. Panel B shows participants raw proportion correct at each m-afc condition (top) as well as the parameter estimates obtained from fitting the Gaussian SDT model (middle) and softmax model (bottom) separately to the percent correct from each m-afc condition. In each figure, the black dot and error bar denote the average and standard error of the mean across participants within each condition. The black dotted line in Panel B, denotes the mean across participants and condition. The fact that estimates of d' are more stable than estimates of β across m-afc conditions is consistent with the model comparison favoring Gaussian SD.

both models, memory performance did not decrease systematically with an increase in the number of alternatives.

Model flexibility

Based on the LL, both models yielded identical fits to the data ($\bar{X} = -105.4$; $SEM = 3.6$ for both models; $t(29) = 0$). Again, this indicates that the superior performance of the Gaussian signal detection model is not due to its having a more flexible functional form.

Critical test for parameter invariance

Panel B in Figure 5 shows the difference in log likelihood (LL) between the Gaussian signal detection and softmax model when parameters d' and β, respectively are fixed across m-afc conditions. As before, positive and negative values indicate support for the Gaussian signal detection and softmax model, respectively, whereas values near zero indicate equal support for both models. We found that the LL was significantly higher for the Gaussian signal detection ($\bar{X} = -106.9$; $SEM = 19$) than the softmax ($\bar{X} = -107.8$; $SEM = 20$) model ($t(29) = 4.42$, $p < .001$; $d_z = .81$). Average parameter estimates in the fixed models were $d' = 1.52$ ($SEM = .12$) and $\beta = 2.04$ ($SEM = .16$). Based on the standard deviation of parameters across m-afc conditions, we also found that there was significantly less variability in parameters in the Gaussian signal detection ($\bar{X} = .24$; $SEM =$

.03) than the softmax model $\bar{X} = .41$; $SEM = .04$) when these were allowed to vary freely ($t(29) = 8.77$, $p < .001$; $d_z = 1.60$). Once again, these results provide support for the Gaussian signal detection over the softmax model.

General Discussion

We revisited the connection between the Gaussian signal detection and (Luce choice) softmax model. Although these two models come from different traditions, they closely approximate each other in detection tasks, and both can be recast as different parametric variants of the signal detection model in the *m*-afc task. Thus, the distinction between signal detection and softmax choice models can be understood as embodying different assumptions about the latent distribution of decision variables in a signal detection model, where the Gaussian parameterization is consistent with pooling of sensory evidence to compute decision variables, whereas the Gumbel distribution is most consistent with taking the maximum of sensory evidence to compute decision variables (Thompson & Singh, 1967). Together, comparing these models may help elucidate how people compute decision variables from sensory evidence in a range of cognitive tasks.

We applied these ideas to examine which signal detection model provides the best characterization of processes in visual working memory tasks. To this end, we designed a critical test to assess which model's parameters are invariant across changes in the number of alternatives in *m*-afc visual memory tasks. We assumed that the model that best capture stable latent processes, should yield parameters that are invariant across *m*-afc conditions (Busemeyer & Wang, 2000) because the computations people use to compute decision variables in these conditions should be the same. We implemented this test in two different visual memory experiments, where we varied the structure of the stimulus space – that is, how similar stimuli were to one another, the dimensionality of stimuli – that is, whether people had to remember simple features (color) or complex real-world objects, and presentation format – that is, whether stimuli were presented simultaneously or sequentially. Across these experiments, we found consistent support for the Gaussian signal detection model. These results align with the view that sensory evidence is pooled via summation or averaging, and indicates that out of this suite of models, the Gaussian signal detection model best capture latent processes in visual memory.

Models of visual working memory

Our work has direct implications for contemporary models of visual memory. First, this is relevant for building cognitive models of visual memory. Relevant in this context are two prominent models, the Target Confusability and Competition (TCC) (Schurgin et al., 2020) and Interference model (Oberauer & Lin, 2017). The TCC model combines principles from signal detection theory and Shepard's law of generalization; it postulates that familiarity is a function of the psychophysical similarity to remembered items, such that items that are more similar to items held in memory generate a stronger familiarity signal. Importantly, an assumption of this model is that the response function that maps familiarity signals to responses is a Gaussian signal detection model. The interference model postulates that memory for items is driven by cued based retrieval. More precisely, access to working memory representations is determined by a spatial retrieval cue, as well as noise that is uniformly distributed across memoranda. In contrast to TCC, the Interference model uses a softmax response function. Importantly, when proposing these models, these researchers did not provide a process-based justification for using one response function over the other.

Our work suggests that the Gaussian signal detection model is more appropriate because it does a superior job of capturing cognitive invariants in forced choice memory tasks. As discussed, this result suggests that people pool sensory evidence via summation or averaging when computing decision variables. Critically, this study is one of few to directly model how people translate early sensory signals to higher-level representations, and lays the groundwork for building and constraining cognitive architectures that characterize the linking function between perception and memory (e.g., Hedayati et al., 2022).

Limitations and future directions

Throughout our article we focused on two specific models of choice the Gaussian signal detection and softmax model. In principle, however, we could have compared a much wider range of models; for instance, we could have considered a larger range of signal detection models with different parameterizations. This approach was taken by Rouder et al., 2010, who used Receiver Operating Characteristics analysis to compare different parameterizations of the signal detection model to a variant of the Gaussian signal detection model, which is most prominent in the recognition memory domain (Wixted, 2007). For instance, the authors considered signal detection models with a log-normal and gamma parameterization. In the current study, we focused on comparing Gaussian signal detection and softmax (Gumbel signal detection) models because they are prominent across different research domains. Furthermore, there is a large body of classic work that examines the formal relationship between these models, but it is disconnected from more contemporary modeling of visual memory. Another major reason is that, unlike the Gaussian and Gumbel signal detection models, these alternative parameterizations do not currently have a clear theoretical interpretation. In short, there is an extremely wide range of possible parameterizations of signal detection models. Considering a larger subset of these is outside of the scope of the current project because our goal is to focus

solely on theoretically-motivated models.

Finally, our results conflict with a recent analysis by Oberauer, 2021. Oberauer, 2021 implemented a factorial comparison of visual working memory models, and found support for the softmax over the Gaussian signal detection model. A major limitation of this work, is that it is not based on a critical test, such as our test of parameter invariance. Instead, Oberauer, 2021 factorially combined different dimensions of each model until he identified a model that provided superior "fit" to the data based on a particular model comparison technique. More precisely, Oberauer considered different combinations of activation functions (e.g., Laplace versus von-Mises) and response rules (e.g., Gaussian signal detection versus softmax), and found that the best fitting model had a von Mises activation and softmax response rule. A critical limitation of this work, is that models were evaluated solely on their fit to data, rather than their ability to capture cognitive invariants. It is known that superior fit to data alone does not entail that a model's basis theory is also a superior one (Roberts & Pashler, 2000). Instead, it could reflect other factors such as, inadequate penalization of a model's flexibility (Piantadosi, 2018; Pitt & Myung, 2002). Our results suggest that the Gaussian signal detection model performs well across a range of experimental conditions and when we make minimal assumptions about the latent activation function.

Conclusion

We revisited the connection between the Gaussian signal detection and Luce choice-based softmax model. We found that the Gaussian signal detection model best captures decision processes that underpin mainstream visual working memory tasks. This result suggest that people pool sensory evidence to compute decision variables in such tasks, and paves the way for developing linking propositions (Teller, 1984) between neural and cognitive models of visual memory.

References

Averbeck, B. B., Latham, P. E., & Pouget, A. (2006). Neural correlations, population coding and computation. *Nature reviews neuroscience*, *7*(5), 358–366.

Balakrishnan, J. (1999). Decision processes in discrimination: Fundamental misrepresentations of signal detection theory. *Journal of Experimental Psychology: Human Perception and Performance*, *25*(5), 1189.

Bates, C. J., Alvarez, G., & Gershman, S. J. (2023). Scaling models of visual working memory to natural images. *bioRxiv*, 2023–03.

Benjamin, A. S., Diaz, M., & Wee, S. (2009). Signal detection with criterion noise: Applications to recognition memory. *Psychological review*, *116*(1), 84.

Bradley, R. A., & Terry, M. E. (1952). Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, *39*(3/4), 324–345.

Brady, T. F., Konkle, T., Alvarez, G. A., & Oliva, A. (2008). Visual long-term memory has a massive storage capacity for object details. *Proceedings of the National Academy of Sciences*, *105*(38), 14325–14329.

Brady, T. F., & Störmer, V. (2020). Greater capacity for objects than colors in visual working memory: Comparing memory across stimulus spaces requires maximally dissimilar foils.

Bridle, J. S. (1990). Probabilistic interpretation of feedforward classification network outputs, with relationships to statistical pattern recognition. *Neurocomputing* (pp. 227–236). Springer.

Busemeyer, J. R., & Wang, Y.-M. (2000). Model comparisons and model selections based on generalization criterion methodology. *Journal of Mathematical Psychology*, *44*(1), 171–189.

Clark, S. E., & Gronlund, S. D. (1996). Global matching models of recognition memory: How the models match the data. *Psychonomic bulletin & review*, *3*(1), 37–60.

Dosher, B. A., & Lu, Z.-L. (1998). Perceptual learning reflects external noise filtering and internal noise reduction through channel reweighting. *Proceedings of the National Academy of Sciences*, *95*(23), 13988–13993.

Falmagne, J.-C., & Doble, C. (2015). *On meaningful scientific laws*. Springer.

Green, D. M., Swets, J. A. et al. (1966). *Signal detection theory and psychophysics* (Vol. 1). Wiley New York.

Gumbel, E. J. (1954). *Statistical theory of extreme values and some practical applications: A series of lectures* (Vol. 33). US Government Printing Office.

Hedayati, S., O'Donnell, R. E., & Wyble, B. (2022). A model of working memory for latent representations. *Nature Human Behaviour*, *6*(5), 709–719.

Holman, E., & Marley, A. (1974). Stimulus and response measurement. *Psychophysical Judgment and Measurement*, *2*, 173.

Kellen, D., Winiger, S., Dunn, J. C., & Singmann, H. (2021). Testing the foundations of signal detection theory in recognition memory. *Psychological review*, *128*(6), 1022.

Krantz, D. H., Luce, R. D., Suppes, P., & Tver-sky, A. (1971). Foundations of measurement: Addıtive and polynomial. *Representations*, *1*.

Luce, R. D. (1959). Individual choice behavior: A theoretical analysis, new york, ny: John willey and sons.

McFadden, D. (1980). Econometric models for probabilistic choice among products. *Journal of Business*, S13–S29.

Mueller, S. T., & Weidemann, C. T. (2008). Decision noise: An explanation for observed violations of signal detection theory. *Psychonomic bulletin & review*, *15*, 465–494.

Murphy, K. P. (2012). *Machine learning: A probabilistic perspective*. MIT press.

Myung, I. J., & Pitt, M. A. (1997). Applying occam's razor in modeling cognition: A bayesian approach. *Psychonomic bulletin & review*, *4*, 79–95.

Navarro, D. J. (2021). If mathematical psychology did not exist we might need to invent it: A comment on theory building in psychology. *Perspectives on Psychological Science*, *16*(4), 707–716.

Oberauer, K. (2021). Measurement models for visual working memory—a factorial model comparison. *Psychological review*.

Oberauer, K., & Lin, H.-Y. (2017). An interference model of visual working memory. *Psychological review*, *124*(1), 21.

Piantadosi, S. T. (2018). One parameter is always enough. *AIP Advances*, *8*(9), 095118.

Pitt, M. A., & Myung, I. J. (2002). When a good fit can be bad. *Trends in cognitive sciences*, *6*(10), 421–425.

Pitt, M. A., Myung, I. J., & Zhang, S. (2002). Toward a method of selecting among computational models of cognition. *Psychological review*, *109*(3), 472.

Pleskac, T. J. (2015). Decision and choice: Luce's choice axiom. *International encyclopedia of the social & behavioral sciences*, *5*, 895–900.

Roberts, S., & Pashler, H. (2000). How persuasive is a good fit? a comment on theory testing. *Psychological review*, *107*(2), 358.

Robinson, M. M., Benjamin, A. S., & Irwin, D. E. (2020). Is there a k in capacity? assessing the structure of

visual short-term memory. *Cognitive Psychology*, *121*, 101305.

Robinson, M. M., Williams, J. R., & Brady, T. F. (2022). What does it take to falsify a psychological theory? a case study on recognition models of visual working-memory.

Rouder, J. N., Morey, R. D., Cowan, N., Zwilling, C. E., Morey, C. C., & Pratte, M. S. (2008). An assessment of fixed-capacity models of visual working memory. *Proceedings of the National Academy of Sciences*, *105*(16), 5975–5979.

Rouder, J. N., Pratte, M. S., & Morey, R. D. (2010). Latent mnemonic strengths are latent: A comment on mickes, wixted, and wais (2007). *Psychonomic Bulletin Review*, *17*, 427–435.

Schurgin, M. W., Wixted, J. T., & Brady, T. F. (2020). Psychophysical scaling reveals a unified theory of visual memory strength. *Nature human behaviour*, *4*(11), 1156–1172.

Suppes, P., & Krantz, D. H. (2007). *Foundations of measurement: Geometrical, threshold, and probabilistic representations* (Vol. 2). Courier Corporation.

Sutton, R. S., & Barto, A. G. (2018). *Reinforcement learning: An introduction*. MIT press.

Swets, J. A. (1959). Indices of signal detectability obtained with various psychophysical procedures. *The Journal of the Acoustical Society of America*, *31*(4), 511–513.

Teller, D. Y. (1984). Linking propositions. *Vision research*, *24*(10), 1233–1246.

Thompson, W., & Singh, J. (1967). The use of limit theorems in paired comparison model building. *Psychometrika*, *32*(3), 255–264.

Thurstone, L. L. (1927). A law of comparative judgment. *Psychological review*, *34*(4), 273.

Townsend, J. T., & Ashby, F. G. (1982). Experimental test of contemporary mathematical models of visual letter recognition. *Journal of Experimental Psychology: Human Perception and Performance*, *8*(6), 834.

Townsend, J. T., & Landon, D. E. (1983). Mathematical models of recognition and confusion in psychology. *Mathematical Social Sciences*, *4*(1), 25–71.

Treisman, M., & Faulkner, A. (1985). On the choice between choice theory and signal detection theory. *The Quarterly Journal of Experimental Psychology*, *37*(3), 387–405.

Wickens, T. D. (2001). *Elementary signal detection theory*. Oxford university press.

Williams, J. R., Robinson, M. M., Schurgin, M., Wixted, J., & Brady, T. F. (2022). You can't "count" how many items people remember in working memory: The importance of signal detection-based measures for understanding change detection performance.

Winiger, S., Singmann, H., & Kellen, D. (2021). Bias in confidence: A critical test for discrete-state models of change detection. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *47*(3), 387.

Wixted, J. T. (2007). Dual-process theory and signal-detection theory of recognition memory. *Psychological review*, *114*(1), 152.

Wixted, J. T. (2020). The forgotten history of signal detection theory. *Journal of experimental psychology: learning, memory, and cognition*, *46*(2), 201.

Yellott Jr, J. I. (1977). The relationship between luce's choice axiom, thurstone's theory of comparative judgment, and the double exponential distribution. *Journal of Mathematical Psychology*, *15*(2), 109–144.

Appendix

Gumbel signal detection and softmax model

We provide a proof for equivalence between the Gumbel signal detection and softmax model (original result proved by Holman & Marley, 1974; Yellott Jr, 1977). We start with the general form of the signal detection model likelihood for discrimination tasks, given partially in the main text in Equation 3. For simplicity, first consider n independent and identically distributed (i.i.d.) variables with probability and cumulative densities f and F, respectively. Equation 3 can be rewritten as a likelihood of the signal detection model for discrimination tasks in terms of these densities of n variables, as shown below

$$P(ID(i)) = P(\forall j \neq i : X_i > X_j), \tag{A1}$$
$$= P(X_i = x)P(X_j < x, \forall j \neq i), \tag{A2}$$

Equation A2 shows the joint probability that the magnitude of the decision variable X_i exceeds the magnitude of all remaining variables X_j, where $\forall j \neq i$.

For continuous variables in a memory task where there is one target and $n-1$ foils, the likelihood that the target generates the maximum familiarity signal is

$$\int_{-\infty}^{\infty} f_T(x)F_F(x)^{n-1}dx, \tag{A3}$$

and the likelihood that a foil generates the maximum familiarity signal is

$$\int_{-\infty}^{\infty} (n-1)f_F(x)F_T(x)F_F(x)^{n-2}dx. \tag{A4}$$

Informally, Equation A3 gives the probability that the target generates a familiarity signal x (denoted with probability density f_T), which exceeds the familiarity signal of all $n-1$ foils (denoted with cumulative density F_T exponentiated by $n-1$). Similarly, Equation A4 gives the probability that one of the foils generates a familiarity signal that exceeds the target and the remaining $n-2$ foils, which can happen in $n-1$ ways. In both equations, these probabilities are integrated over every possible value of x.

Next, assume that each of n variables has a Gumbel distribution (for maximums) with scale parameter $\alpha = 1$. As before, we assume that on target present trials decision variables will be larger on average than on target absent trials, so the shift parameter $\mu > 0$ and $\mu = 0$ on target present and target absent trials, respectively. Thus, the densities for decision variables elicited by the target, f_T and F_T on target present trials are

$$f_T(x) = e^{\mu - x}e^{-e^{\mu - x}}, \tag{A5}$$
$$F_T(x) = e^{-e^{\mu - x}}, \tag{A6}$$

and the densities for decision variables elicited by the foils, f_F and F_F on target absents trials are

$$f_F(x) = e^{-x}e^{-e^{-x}}, \tag{A7}$$
$$F_F(x) = e^{-e^{-x}}. \tag{A8}$$

Replacing the generic densities in Equation A4 with the Gumbel densities in Equations A5 through A8, the likelihood for the Gumbel signal detection model on target present and absent trials is the following,

$$P(ID(\text{Target})) = \int_{-\infty}^{\infty} (e^{\mu - x}e^{-e^{\mu - x}})(e^{-e^{-x}})^{n-1}dx, \tag{A9}$$

$$P(ID(\text{Foil})) = \int_{-\infty}^{\infty} (n-1)(e^{-x}e^{-e^{-x}})(e^{-e^{\mu - x}})(e^{-e^{-x}})^{n-2}dx. \tag{A10}$$

For simplicity, we show equivalence between the Gumbel signal detection and softmax model using the likelihood for target present trials (Equation A9), but these steps can be extended to the likelihood for target absent trials (Equation A10).

First, using substitution, set $z = e^{-e^{-x}}$. Differentiating, $\frac{dz}{dx} = e^{-e^{-x}-x}$ and $dx = (e^{-e^{-x}-x})^{-1}dz$. Simplifying,

$$P(ID(\text{Target})) = \int_{-\infty}^{\infty} (e^{\mu - x}e^{-e^{\mu - x}})z^{n-1}(e^{-e^{-x}-x})^{-1}dz \tag{A11}$$
$$= \int_{-\infty}^{\infty} e^{\mu}e^{-e^{\mu - x}+e^{-x}}z^{n-1}dz. \tag{A12}$$

Replacing x with $-ln(-ln(z))$, in Equation A12 and simplifying gives,

$$P(ID(\text{Target})) == \int_{-\infty}^{\infty} e^{\mu}e^{-e^{ln(-ln(z))+\mu}+e^{ln(-ln(z))}}z^{n-1}dz, \tag{A13}$$
$$= \int_{-\infty}^{\infty} e^{\mu}e^{ln(z)e^{\mu}-ln(z)}z^{n-1}dz, \tag{A14}$$
$$= \int_{-\infty}^{\infty} e^{\mu}z^{e^{\mu}}z^{-1}z^{n-1}dz. \tag{A15}$$

After applying the power rule, Equation A15 can be rewritten as,

$$P(ID(\text{Target})) = \int_{-\infty}^{\infty} e^{\mu}z^{e^{\mu}+n-2}dz = e^{\mu}\frac{z^{e^{\mu}+n-1}}{e^{\mu}+n-1}. \tag{A16}$$

Substituting $e^{-e^{-x}}$ back for z, and plugging in the boundaries, yields

$$P(ID(\text{Target})) = \frac{e^{\mu - e^{-x}(e^{\mu}+n-1)}}{e^{\mu}+n-1}\Big|_{-\infty}^{\infty}, \tag{A17}$$

$$= \frac{e^{\mu}}{e^{\mu}+n-1}. \tag{A18}$$

Equation A18 is identical to softmax expression for $P(ID(\text{Target}))$ in discrimination tasks (Equation 14 in main text) with $\beta = \mu$, completing the proof.